大数据分析得天下　大数据应用赢未来

大数据

技术及应用教程

李联宁　编著



清华大学出版社

大数据技术及应用教程

李联宁　编著

清华大学出版社
北　京

内 容 简 介

本书详细介绍了大数据技术的基础理论和最新主流前沿技术，全书共分为10章，分别介绍我们目前面临的数字化信息社会的大数据时代、大数据技术基本概念、云计算网络、大数据采集与预处理、大数据存储、计算模式与处理系统、查询显示与交互、大数据分析与数据挖掘、隐私与安全、大数据技术发展前景，同时包括行业案例研究（银行、保险、证券、金融行业），典型系统与相关大数据分析实例。

本书主要作为高等院校计算机专业、信息管理专业、经济类专业、管理类专业相关本科生和研究生专业基础课的教材，也可以作为干部培训、职业技术教育以及职业培训机构的云计算与大数据分析技术的专业训练教材。对从事云计算与大数据分析工作的财政金融、政府管理、计算机网络、软件工程的方面的管理与工程技术人员也有学习参考价值。

图书在版编目（CIP）数据

大数据技术及应用教程/李联宁编著. —北京：清华大学出版社，2016
ISBN 978-7-302-44561-6

Ⅰ. ①大…　Ⅱ. ①李…　Ⅲ. ①数据处理—教材　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2016）第174858号

责任编辑：白立军　李　晔
封面设计：杨玉兰
责任校对：李建庄
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课 件 下 载：http://www.tup.com.cn，010-62795954
印 装 者：北京密云胶印厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：23.75　　**字　　数**：576千字
版　　次：2016年10月第1版　　**印　　次**：2016年10月第1次印刷
印　　数：1～2000
定　　价：49.00元

产品编号：069155-01

前言

本书试图在介绍大数据技术的理论基础上对大数据分析最新前沿技术做全面详细介绍,给出实际案例及行业解决方案,达到技术全面、案例教学及工程实用的目的。

本书主要分为4个部分,共10章,分别按大数据的技术架构分层次详细讲述涉及大数据分析系统的各类相关技术:

第一部分　大数据基础知识,简单介绍我们目前面临的数字化时代与信息社会的状况,大数据的定义和特点、大数据技术基础、大数据的社会价值、大数据的商业应用、大数据的基础架构、云计算网络的技术层次、典型的云计算网络平台,包括第1章“大数据技术基本概念”和第2章“基础架构——云计算网络”;

第二部分　大数据理论与技术,介绍涉及大数据分析的基本理论与技术基础,按照技术层次分别介绍大数据采集与预处理、大数据存储、大数据计算模式与处理系统、大数据查询、显示与交互、大数据分析与数据挖掘、大数据隐私与安全,包括第3章到第8章的内容;

第三部分　为行业案例研究,以银行、保险、证券、金融行业为例,介绍涉及大数据分析的理论与技术方法在具体行业中的应用,包括第9章“行业案例研究”;

第四部分　大数据技术发展前景,介绍大数据引发的新一代信息技术变革浪潮、大数据各个过程的最新技术与发展前景,包括第10章大数据技术发展前景。

本书主要作为高等院校计算机专业、信息管理与信息系统专业、经济类专业、管理类专业相关专业本科生和研究生专业基础课的教材,安排课时为48课时(3学分)。如课时缩减,可在概要叙述第一部分的基础上,主要讲解第二部分第3章到第8章的内容,并安排学生在课外自主阅读每章节后的案例及第9章“行业案例研究”。第10章“大数据技术发展前景”仅作参考性讲解。

本书的特点是紧扣实践应用需求,全面讲述讲云计算与大数据分析实用技术,提供了大量的实际案例、数据分析适用技术。内容新颖、用表格和结构图直观描述知识并力图反映最新主流技术。

每一章在讲解相关理论外,还讲解了最新前沿技术。各章都附有案例、习题以帮助读者学习理解和实际工程应用。为方便教师教学,附有全套教学PPT课件、教学大纲、教学计划以便教师使用。

本书由李联宁教授编著,在本书编写过程中,编者参考了国内外大量的云计算网络与大数据分析技术的书刊及文献资料,主要参考书籍及研究论文在书后“参考文献”中

列出。但由于大量来自网络的资料未能详尽标注作者及文献资料来源,疏漏之处在所难免,在此一并对书刊文献、科技论文的作者表示感谢。如有遗漏,恳请相应书刊文献作者及时告知,将在书籍再版时列入。如发现本书有错误或不妥之处,恳请广大读者不吝赐教。

编　者

2016 年 8 月

目录

第一部分 大数据基础

第二部分 大数据技术

第三部分　大数据分析案例

第四部分　大数据技术现状及发展展望



大数据基础

第 1 章　大数据技术基本概念

当今，信息技术为人类步入智能社会开启了大门，带动了互联网、物联网、电子商务、现代物流、网络金融等现代服务业发展，催生了车联网、智能电网、新能源、智能交通、智能城市、高端装备制造等新兴产业发展。现代信息技术正成为各行各业运营和发展的引擎。但这个引擎正面临着大数据这个巨大的考验。各种业务数据正以几何级数的形式爆发，其格式、收集、储存、检索、分析、应用等诸多问题，不再能以传统的信息处理技术加以解决，对人类实现数字社会、网络社会和智能社会带来了极大的障碍。

大数据的出现将影响各行各业以及每个人生活。以下十个事实会让你相信，每个人都必须注意大数据：

(1) 全球数据的 90%产生于过去 2 年内。

(2) 当前数据产生的速度非常快，以今天的数据生产速度，我们可以在 2 天内生产出 2003 年以前的所有数据。

(3) 行业内获取并且存储的数据量每 1.2 年就会翻一番。

(4) 到 2020 年，全球数据量将由现在的 3.2ZB 变为 40ZB(1ZB＝1024EB，1EB＝1024PB，1PB＝1024TB)。

(5) 仅 Google 一家搜索引擎，每秒就处理 4 万次搜索查询，一天之内更是超过 35 亿次。

(6) 最近的统计报告显示，我们每分钟在 Facebook 上贡献 180 万次赞，上传 20 万张照片。与此同时，我们每分钟还发送 2.04 亿封邮件，发送 27.8 万个推文。

(7) 每分钟大约有 100 小时的视频被传上类似 YouTube 这样的视频网站。更有趣的是，要花费 15 年才能看完一天之内被传到 YouTube 上的全部视频。

(8) AT&T 被认为是能够用单一数据库存储最多数据量的数据中心。

(9) 在美国，很多新的 IT 工作将被创造出来以处理即将到来的大数据工程潮，而每个这样的职位都将需要 3 个额外职位的支持，这将会带来总计 600 万个新增工作岗位。

(10) 全球每分钟会新增 570 个网站。这一统计数字至关重要，也具有颠覆性。

预测是：数据以及数据分析能力正与日俱增，未来五年，无论何等规模的企业都将使用某种形式的数据分析来影响其商业运作。

1.1　数据

数据(data)是对客观事物的逻辑归纳，用符号、字母等方式对客观事物进行直观描述。数据是进行各种统计、计算、科学研究或技术设计等所依据的数值，是表达知识的字符的集合。数据是信息的表现形式。数据可以是连续的值，例如声音，称为模拟数据；也

可以是不连续(离散)的值,例如成绩,称为数字数据。

1.1.1 数据的单位

数据最小的基本单位是 bit,按顺序给出所有单位:bit、Byte、KB、MB、GB、TB、PB、EB、ZB、YB、BB、NB、DB。

它们按照进率 1024(2 的十次方)来计算:

1Byte = 8bit
1KB = 1024Bytes = 8192bit
1MB = 1024KB = 1 048 576Bytes
1GB = 1024MB = 1 048 576KB
1TB = 1024GB = 1 048 576MB
1PB = 1024TB = 1 048 576GB
1EB = 1024PB = 1 048 576TB
1ZB = 1024EB = 1 048 576PB
1YB = 1024ZB = 1 048 576EB
1BB = 1024YB = 1 048 576ZB
1NB = 1024BB = 1 048 576YB
1DB = 1024NB = 1 048 576BB

1.1.2 数据与信息的关系

数据是一种未经加工的原始资料。数字、文字、符号、图像都是数据。数据是客观对象的表示,而信息则是数据内涵的意义,是数据的内容和解释。综上所述,数据就是指能够客观反映事实的数字和资料。

信息与数据的关系是:信息与数据是不可分离的,数据是信息的表达,信息是数据的内涵。数据本身并没有意义数据只有对实体行为产生影响时才成为信息。

1.1.3 数据的分类

在信息社会,信息可以划分为两大类:一类信息能够用数据或统一的结构加以表示,我们称之为结构化数据,如数字、符号;另一类信息无法用数字或统一的结构表示,如文本、图像、声音、网页等,我们称之为非结构化数据。结构化数据属于非结构化数据的一部分,是非结构化数据的特例。

1. 结构化数据

结构化信息是指信息经过分析后可分解成多个互相关联的组成部分,各组成部分间有明确的层次结构,其使用和维护通过数据库进行管理,并有一定的操作规范。我们通常接触的,包括生产、业务、交易、客户信息等方面的记录都属于结构化信息。

结构化数据简单来说就是存储在结构化数据库里的数据,可以用二维表结构来逻辑表达实现的数据。结合到典型场景中更容易理解,比如企业 ERP、财务系统;医疗 HIS 数

据库;教育一卡通;政府行政审批;其他核心数据库等。这些应用需要包括高速存储应用需求、数据备份需求、数据共享需求以及数据容灾需求。

2. 非结构化数据

不方便用数据库二维逻辑表来表现的数据即称为非结构化数据,包括所有格式的办公文档、文本、图片、标准通用标记语言下的子集 XML、HTML、各类报表、图像和音频/视频信息等等。

所谓非结构化数据库,是指数据库的变长记录由若干不可重复和可重复的字段组成,而每个字段又可由若干不可重复和可重复的子字段组成。用它不仅可以处理结构化数据(如数字、符号等信息)而且更适合处理非结构化数据(全文文本、图像、声音、影视、超媒体等信息)。简单地说,非结构化数据库就是字段可变的数据库。

非结构化 Web 数据库主要是针对非结构化数据而产生的,与以往流行的关系数据库相比,其最大区别在于它突破了关系数据库结构定义不易改变和数据定长的限制,支持重复字段、子字段以及变长字段并实现了对变长数据和重复字段进行处理和数据项的变长存储管理,在处理连续信息(包括全文信息)和非结构化信息(包括各种多媒体信息)中有着传统关系型数据库所无法比拟的优势。

3. 半结构化数据

所谓半结构化数据,就是介于完全结构化数据(如关系型数据库、面向对象数据库中的数据)和完全无结构的数据(如声音、图像文件等)之间的数据,HTML 文档就属于半结构化数据。它一般是自描述的,数据的结构和内容混在一起,没有明显的区分。

4. 各类数据的区别

结构化数据：行数据,存储在数据库里,可以用二维表结构来逻辑表达实现的数据。

非结构化数据：包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频/视频信息等等。

半结构化数据：介于完全结构化数据和完全无结构的数据之间的数据,它一般是自描述的,数据的结构和内容混在一起。

1）数据模型

各类数据的数据模型和基本特征如下：

结构化数据：二维表(关系型)。

半结构化数据：树、图。

非结构化数据：无。

2）关系型数据库系统 RMDBS 的数据模型

RMDBS 的数据模型包括网状数据模型、层次数据模型、关系型。

3）不同类型数据的形成过程

结构化数据：先有结构,再有数据。

半结构化数据：先有数据,再有结构。

5. 互联网信息分类

互联网上出现的海量信息,同样分为结构化、半结构化和非结构化三种。

(1) 结构化信息如电子商务信息,信息的性质和量值的出现的位置是固定的,如图 1.1 所示;

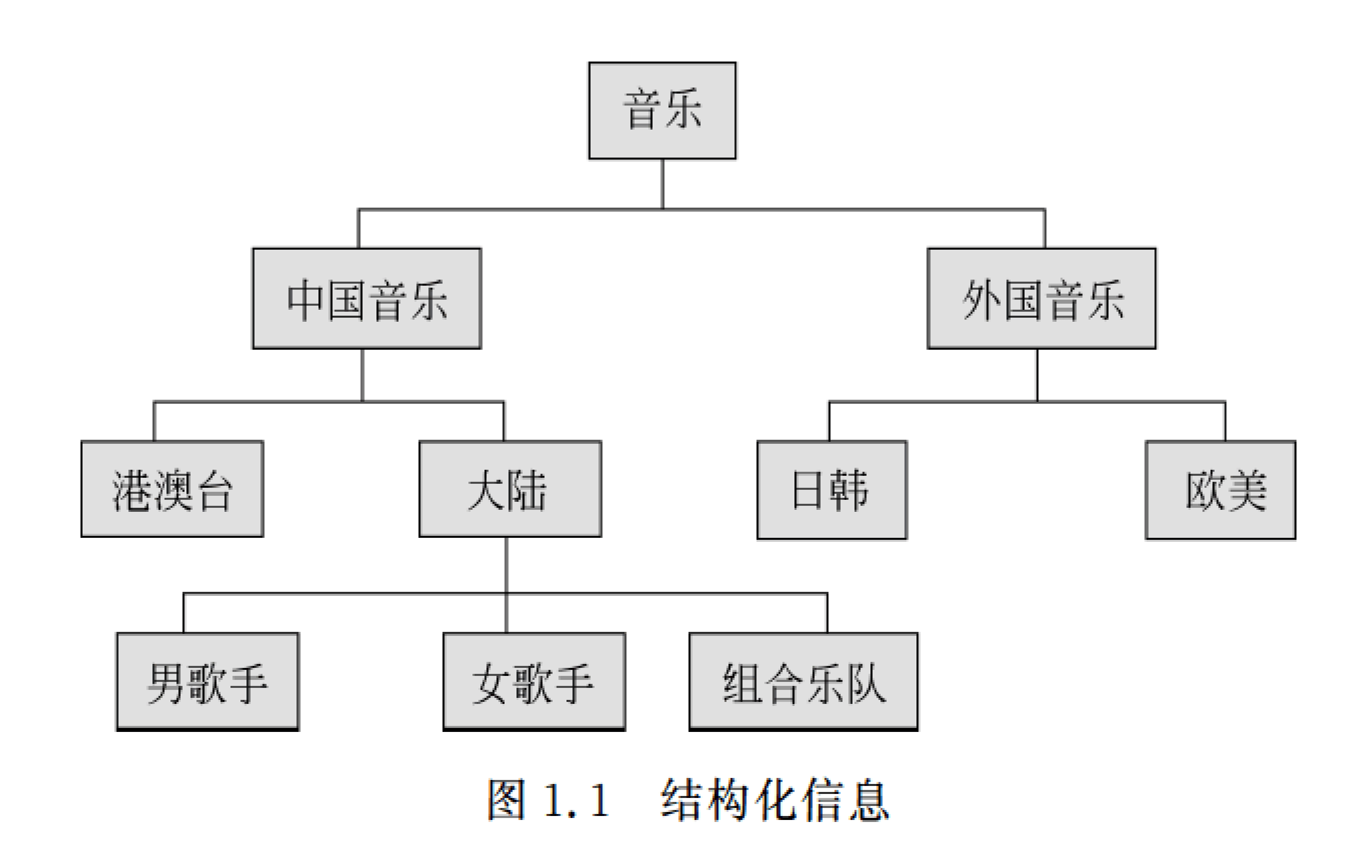


图 1.1 结构化信息

(2) 半结构化的信息如专业网站上的细分频道,其标题和正文的语法相当规范,关键词的范围相当局限;

(3) 非结构化的信息如博客(BLOG)和网上社区 BBS,所有内容都是不可预知的。

结构化信息和非结构化信息是 IT 应用的两个世界,它们有着各自不同的应用进化特点和规律。但是,这两个世界之间还缺少相互连接的桥梁,而这种缺失使企业中不可避免地存在"活动""信息和知识"的分离,其后果就是:虽然它们都在进行着"知识化"的努力,但两个世界分离的 IT 应用模式,注定使其难以真正实现它们的初衷——"在最合适的时间,将最合适的信息传送给最合适的人"。

6. 中国企业的数据现状

目前,中国企业 500 强的每日数据生成量近一半都多于 1GB,更有 4.9%的企业超过 1TB。中国企业级数据中心数据存储量正在快速增长,非结构化数据呈指数倍增长,如果能有效地处理和分析,非结构数据中也富含对企业非常有价值的信息,如图 1.2 所示。

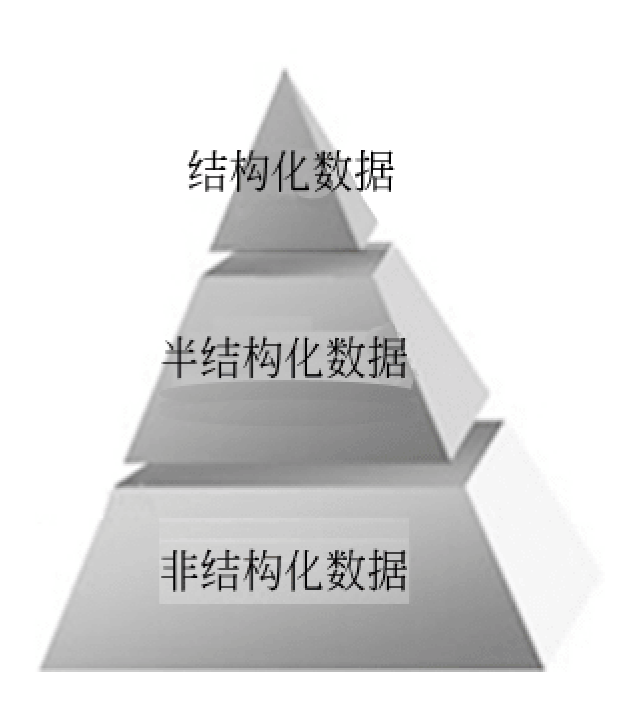


图 1.2 中国企业的数据现状

1.2 信息

1.2.1 信息的定义

"信息"一词在英文、法文、德文、西班牙文中均是 information,日文中为"情报",我国台湾称之为"资讯",我国古代用的是"消息"。

信息,指音讯、消息、通信系统传输和处理的对象,泛指人类社会传播的一切内容。人通过获得、识别自然界和社会的不同信息来区别不同的事物,得以认识和改造世界。在一切通信和控制系统中,信息是一种普遍联系的形式。

根据对信息的研究成果。科学的信息概念可以概括如下:

信息是对客观世界中各种事物的运动状态和变化的反映，是客观事物之间相互联系和相互作用的表征，表现的是客观事物运动状态和变化的实质内容。

信息技术是指有关信息的收集、识别、提取、变换、存储、传递、处理、检索、检测、分析和利用等的技术。凡涉及这些过程和技术的工作部门都可称作信息部门。

1.2.2　信息资源

只要事物之间的相互联系和相互作用的存在，就有信息发生。人类社会的一切活动都离不开信息，信息具有使用价值，能够满足人们的特殊需要，可以用来为社会服务。但是，认识到信息是一种独立的资源还是 20 世纪 80 年代以来的事情。

美国哈佛大学的研究小组给出了著名的资源三角形。他们指出：没有物质，什么都不存在；没有能量，什么都不会发生；没有信息，任何事物都没有意义。资源三角形图示如图 1.3 所示。

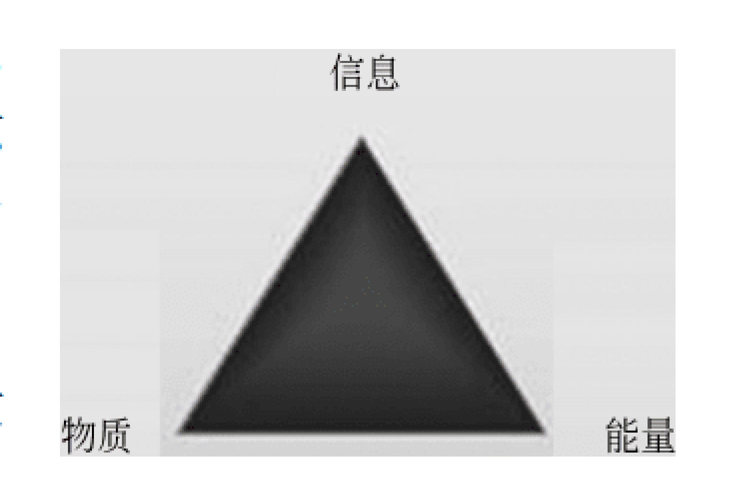


图 1.3　资源三角形

作为资源，物质为人们提供了各种各样的材料；能量提供各种各样的动力；信息提供各种各样的知识。

信息是普遍存在的，但并非所有的信息都是资源。只有满足一定条件的信息才能构成资源。对于信息资源，有狭义和广义之分：

狭义的信息资源，指的是信息本身或信息内容，即经过加工处理，对决策有用的数据。开发利用信息资源的目的就是为了充分发挥信息的效用，实现信息的价值。

广义的信息资源，指的是信息活动中各种要素的总称。“要素”包括信息、信息技术以及相应的设备、资金和人等。

狭义的观点突出了信息是信息资源的核心要素，但忽略了“系统”。事实上，如果只有核心要素，而没有“支持”部分(技术、设备等)，就不能进行有机的配置，不能发挥信息作为资源的最大效用。

归纳起来，可以认为，信息资源由信息生产者、信息、信息技术三大要素组成。

(1) 信息生产者是为了某种目的的生产信息的劳动者，包括原始信息生产者、信息加工者或信息再生产者。

(2) 信息既是信息生产的原料，也是产品。它是信息生产者的劳动成果，对社会各种活动直接产生效用，是信息资源的目标要素。

(3) 信息技术是能够延长或扩展人的信息能力的各种技术的总称，是对声音、图像、文字等数据和各种传感信号的信息进行收集、加工、存储、传递和利用的技术。信息技术作为生产工具，对信息收集、加工、存储和传递提供支持与保障。

1. 特点

信息资源与自然资源、物质资源相比，具有以下几个特点：

(1) 能够重复使用，其价值在使用中得到体现。

(2) 信息资源的利用具有很强的目标导向，不同的信息在不同的用户中体现不同的价值。

(3) 具有整合性。人们对其检索和利用,不受时间、空间、语言、地域和行业的制约。

(4) 它是社会财富,任何人无权全部或永久买下信息的使用权;它是商品,可以被销售、贸易和交换。

(5) 具有流动性。

2. 信息资源作为经济资源的一般特征

(1) 作为生产要素的人类需求性。

(2) 稀缺性:稀缺性是经济资源最基本的经济学特征。

(3) 使用方向的可选择性:关于信息资源的有效配置问题,这是由于信息资源具有很强的渗透性。

3. 与物质资源、能源资源相比,具有一些独有特征

(1) 共享性。

(2) 时效性:只有时机适宜,才能发挥效益。

(3) 动态性:信息资源是一种动态资源,呈现不断丰富、不断增长的趋势。

(4) 不可分性:信息的不可分性表现在它在生产过程中的不可分。

(5) 不同一性:作为资源的信息必定是完全不同一的。

(6) 支配性(即驾驭性):支配性是指信息资源具有开发和支配其他资源的能力。

1.2.3 信息的应用意义

如果说结构化信息更多地忠实、详实地记录了企业的生产交易活动,是显性的表示,那么非结构化信息则隐性包含了掌握着企业命脉的关键,隐含着许多提高企业效益的机会。对于企业来说,企业内部,以及企业与供应商、客户、合作伙伴和员工数字化共享所有形式的数据资源,已越来越重要。

90%的信息和知识在“结构化”世界之外,IT 应用中还存在着一个“非结构化”的世界。对大多数企业来说,ERP 等业务系统所管理的结构化数据只占到企业全部信息和知识的 10%左右,其他的 90%都是数据库难以存取到的非结构化信息和知识。

来自 IDC 的分析显示,虽然很多企业投资不菲建立了诸多业务支撑系统,但仍有 72%的管理者认为知识没有在他们的组织得到重复利用,88%的人认为他们没有接触到企业最佳实践的机会。Gartner 也曾预言,对非结构化信息和知识的管理将会带来一个新 IT 应用潮流。

非结构化信息处理类似于 20 世纪 70 年代以前的结构化信息应用。割裂、无法进行数据互操作的应用是其主流。以人们最常用的文档软件来看,DOC 文档是 Word 的专用格式,WPS、永中、中文 2000 等 Office 产品厂商则各有各的“自留地”。这种情况下,由于文档格式的束缚而使信息四分五裂,信息流无法通畅流转,信息处理更加困难,信息资源因为“信息流的不通畅”而丧失了其应有的巨大价值。

从非结构化到半结构化,从半结构化到结构化,从结构化到关联数据体系,从关联数据体系到数据挖掘,从数据挖掘到故事化呈现,从故事化呈现到决策导向,是信息资源应用的几个不同发展阶段。

1.3　大数据

1.3.1　大数据发展历史

1. 大数据出现的背景

2012 年以来，大数据（big data）一词越来越多地被提及，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。它已经上过《纽约时报》《华尔街日报》的专栏封面，进入美国白宫官网的新闻，现身在国内一些互联网主题的讲座沙龙中，甚至被嗅觉灵敏的证券公司等写进了投资推荐报告。

数据正在迅速膨胀并变大，它决定着企业的未来发展，虽然现在企业可能并没有意识到数据爆炸性增长带来问题的隐患，但是随着时间的推移，人们将越来越多地意识到数据对企业的重要性。大数据时代对人类的数据驾驭能力提出了新的挑战，也为人们获得更为深刻、全面的洞察能力提供了前所未有的空间与潜力。

最早提出大数据时代到来的是全球知名咨询公司麦肯锡，麦肯锡称："数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。""大数据"在物理学、生物学、环境生态学等领域以及军事、金融、通信等行业存在已有时日，却因为近年来互联网和信息行业的发展而引起人们关注。

大数据在互联网行业指的是这样一种现象：互联网公司在日常运营中生成、累积的用户网络行为数据。这些数据的规模是如此庞大，以至于不能用 G 或 T 来衡量，大数据的起始计量单位至少是 P（1000 个 T）、E（100 万个 T）或 Z（10 亿个 T）。

2. 互联网背景下出现的大数据

1）越来越多的私有化的 Web 化数据

电商网站、BBS、知乎问答、互动百科、豆瓣电影等内容便是属于此类。垂直网站在达到一定规模后，拥有与搜索引擎博弈的能力时，便可屏蔽搜索引擎的爬虫，将自己的数据"私有化"。

垂直网站提供的搜索功能，可以用个性化的搜索功能和独有的挖掘能力，提供更好的搜索体验。甚至上升为垂直搜索引擎，如知乎搜索。另外一种垂直搜索引擎即是综合其他垂直的结构化数据，提供搜索服务，如去哪儿、一淘。

随着 Web 的发展，垂直搜索是未来搜索引擎细分的一个方向，且将对传统搜索引擎构成威胁。类似手机上浏览器和原生 APP 之间的关系：浏览器和 APP 流量对半分。我们把传统搜索引擎（如百度）看成这一个浏览器，那么垂直搜索引擎便是 APP。垂直搜索引擎也如 APP 一样正在壮大。且他们具有的核心优势都是：个性化 VS 统一的优势。

如果说 Web 数据私有化使前面提到的"Web 化的信息，能抓取：不能抓取的约为 1∶500"这个比率发生变化。下面要谈的将影响"不到 1%的信息 Web 化"的 1%。

2）巨量增长的没有 Web 化的数据

随着 10 多年的发展，PC 互联网已积累大量的数据；而在移动互联网的浪潮下，APP、

云应用、社交和物联网让数据爆炸式增长。对搜索引擎来说，这些数据几乎都是不可见的。

（1）人工整理的数据。

药监局的数据就是例子。这类数据集中存在于政府部门、机构组织和一些企业手里。他们手里既掌握着民众关心的权威民生数据，又暂时没有将这些数据通过网站开放出来。与此类似的拥有数据的还有交通部门、环保部门、旅游局、卫生局、教育局等民众关注的各个领域。经过十多年的信息化建设，这些数据想必已经达到可观的量级。

另外，“我查查”的条形码数据也可归为此类。我查查团队创业初期，数百人团队在全国商场收集商品条形码数据。我查查有一定规模后，用户才主动为其添加条形码数据。

（2）社交产生的数据。

这里的社交网络不仅仅指微博或人人网。QQ 聊天也是一种社交。邮件也是一种社交。甚至短信通信也是一种社交。我们不妨将这称为“暗社交”。这些社交过程又产生了大量的信息，尤其是分享行为。一定程度上部分社交网站的数据是 Web 化的，但是它们是封闭的。这部分数据正在巨量增长，而搜索引擎对它们无能为力。

（3）APP 产生的数据。

有人曾经抛出过“Web 已死”的说法。移动互联网已经不再是由 Web 通过超链接互相连接的网络。APP 之间通过接口互相链接，APP 上的不同用户通过 QQ 好友关系、微信圈、微博关注关系、手机号码等方式互相链接。而传统搜索引擎正是基于超链接的。带来的实际问题就是，搜索引擎如何搜索啪啪等 APP 的数据？

（4）个人云应用产生的数据。

个人云应用主要是解决多屏同步的问题。这让更多用户选择将数据保存在云端。在不同设备上进行账号认证后下载并使用这些数据。这类应用除了同步通讯录、收藏夹这类私密性强的数据外，还有印象笔记、网易云阅读等类型的大文本数据。个人云应用将越来越多。若干年后，我们认为 Office 提供云同步功能也不是没可能。这些数据，搜索引擎无能为力。

（5）物联网产生的数据。

车联网、监控录像、电子抄表、水文监测等物联网应用每时每刻也在产生大量的数据。这个行业还没爆发。爆发的时候，应用也不会局限于此。互联网链接网页，移动互联网链接天下芸芸众生，而物联网，链接天下万物。现在中国的手机用户数突破 11 亿。芸芸众生基本已连起来。不过相比 11 亿，物联网用户数则是一个惊人的量级。这些“用户”也将产生大量的数据。这些数据将来是否要被人类搜索？以什么形式搜索？搜索的结果是什么？

1.3.2 大数据的定义和特点

信息技术领域原先已经有“海量数据”“大规模数据”等概念，但这些概念只着眼于数据规模本身，未能充分反映数据爆发背景下的数据处理与应用需求，而“大数据”这一新概念不仅指规模庞大的数据对象，也包含对这些数据对象的处理和应用活动，是数据对象、技术与应用三者的统一。

1. 大数据(Big Data,巨量数据集合,IT 行业术语)

大数据或称巨量资料,指的是所涉及的资料量规模巨大到无法通过目前主流软件工具,在合理时间内达到撷取、管理、处理,并整理成为帮助企业经营决策更积极目的的资讯。大数据对象既可能是实际的、有限的数据集合,如某个政府部门或企业掌握的数据库,也可能是虚拟的、无限的数据集合,如微博、微信、社交网络上的全部信息。

在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》中,大数据是指不用随机分析法(抽样调查)这样的捷径,而采用所有数据进行分析处理。

对于“大数据”研究机构 Gartner 给出了这样的定义。“大数据”是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

根据维基百科的定义,“大数据”是一个体量特别大、数据类别特别大的数据集,是指无法在可承受的时间范围内用传统数据库工具对其内容进行抓取、管理和处理的数据集合。

大数据从本质上来讲包含数量、类型、速度 3 个维度的问题,事实上,要想从根本上区别这 3 个维度是不可能的。因为,大数据概念的提出是源于技术的发展。大数据的本质构建如图 1.4 所示。

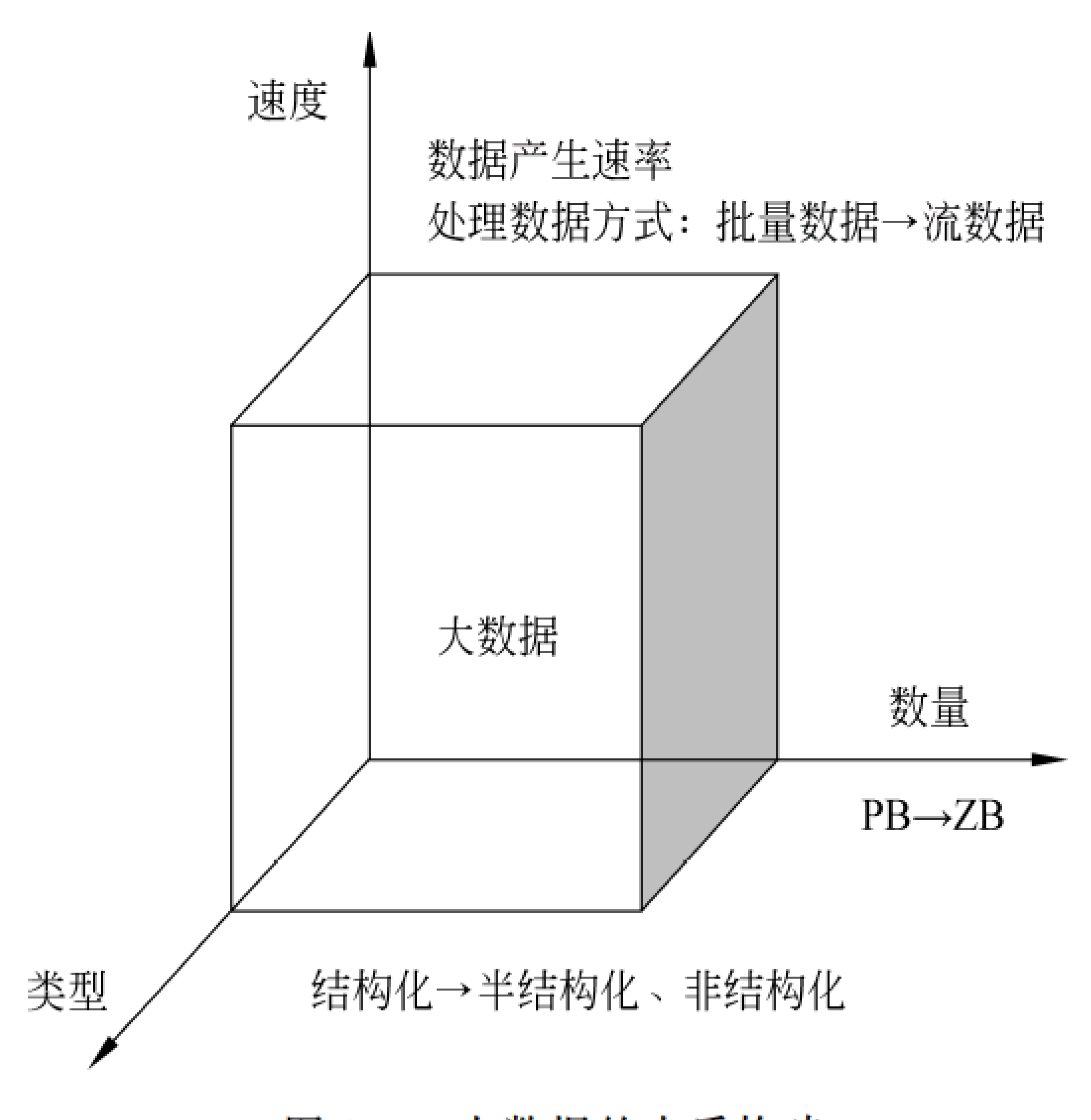


图 1.4　大数据的本质构建

“大数据”首先是指数据体量(Volumes)大,指大型数据集,一般在 10TB 规模左右,但在实际应用中,很多企业用户把多个数据集放在一起,已经形成了 PB 级的数据量。

其次是指数据类别(Variety)大,数据来自多种数据源,数据种类和格式日渐丰富,已冲破了以前所限定的结构化数据范畴,囊括了半结构化和非结构化数据。

接着是数据处理速度(Velocity)快,在数据量非常庞大的情况下,也能够做到数据的实时处理。

最后一个特点是指数据真实性(Veracity)高,随着社交数据、企业内容、交易与应用

数据等新数据源的兴趣，传统数据源的局限被打破，企业愈发需要有效的信息之力以确保其真实性及安全性。

2. 大数据的实质

从狭义的字面含义理解，它应该与小数据相对应，大数据意指数据量特别巨大，超出了我们常规的处理能力，必须引入新的科学工具和技术手段才能够进行处理的数据集合。

所谓的小数据，指的是数据规模比较小，用传统工具和方法足以进行处理的数据集合。比如牛顿时代的各门自然科学，其数据量都不大，第谷观测了20年的天文数据，开普勒很快用手工就处理完毕，并从中发现了开普勒定律。后来，随着科学的发展，数据量有了比较大的增加，为了处理这些当时看来的“大数据”，统计学家创造了抽样方法，由此解决了数据处理难题。

大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。换言之，如果把大数据比作一种产业，那么这种产业实现盈利的关键，在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。

从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台的计算机进行处理，必须采用分布式架构。它的特色在于对海量数据进行分布式数据挖掘，但它必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术。

随着云时代的来临，大数据(Big data)也吸引了越来越多的关注。大数据(Big data)通常用来形容一个公司创造的大量非结构化数据和半结构化数据，这些数据在下载到关系型数据库用于分析时会花费过多的时间和金钱。

大数据分析常和云计算联系到一起，因为实时的大型数据集分析需要像云计算的框架来向数十、数百或甚至数千的计算机分配工作。

大数据需要特殊的技术，以有效地处理大量的可容忍时间内的数据。适用于大数据的技术，包括大规模并行处理(MPP)数据库、数据挖掘、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统。

3. 大数据的特点

业界通常用4个V(即Volume、Variety、Value、Velocity)来概括大数据的特征。具体来说，大数据具有4个基本特征：

第一，Volume(大量)，数据体量巨大，从TB级别，跃升到PB级别。

数据体量(volumes)大，指大型数据集，一般在10TB规模左右，但在实际应用中，很多企业用户把多个数据集放在一起，已经形成了PB级的数据量；百度资料表明，其首页导航每天需要提供的数据超过1.5PB(1PB=1024TB)，这些数据如果打印出来将超过5千亿张A4纸。有资料证实，到目前为止，人类生产的所有印刷材料的数据量仅为200PB。

第二，Variety(多样)，数据类别大和类型多样，即数据类型繁多。除了标准化的结构化编码数据之外，还包括网络日志、视频、图片、地理位置信息等等非结构化或无结构数据。

数据来自多种数据源，数据种类和格式日渐丰富，已冲破了以前所限定的结构化数据范畴，囊括了半结构化和非结构化数据。现在的数据类型不仅是文本形式，更多的是图片、视频、音频、地理位置信息等多类型的数据，个性化数据占绝对多数。

第三，Value（价值），价值真实性高和密度低，即商业价值高，但价值密度低。在数据的海洋中不断寻找，才能“淘”出一些有价值的东西，可谓“沙里淘金”。

随着社交数据、企业内容、交易与应用数据等新数据源的兴起，传统数据源的局限被打破，企业愈发需要有效的信息之力以确保其真实性及安全性。以视频为例，一小时的视频，在不间断的监控过程中，可能有用的数据仅仅只有一两秒。

第四，Velocity（高速），处理速度快，即处理速度快，实时在线。各种数据基本上实时、在线，并能够进行快速的处理、传送和存储，以便全面反映对象的当下状况。

在数据量非常庞大的情况下，也能够做到数据的实时处理。数据处理遵循“1 秒定律”，可从各种类型的数据中快速获得高价值的信息。

有人把数据比喻为蕴藏能量的煤矿。煤炭按照性质有焦煤、无烟煤、肥煤、贫煤等分类，而露天煤矿、深山煤矿的挖掘成本又不一样。与此类似，大数据并不在“大”，而在于“有用”。价值含量、挖掘成本比数量更为重要。对于很多行业而言，如何利用这些大规模数据成为赢得竞争的关键。

大数据的价值体现在以下几个方面：

（1）对大量消费者提供产品或服务的企业可以利用大数据进行精准营销；

（2）做小而美模式的中小型企业可以利用大数据做服务转型；

（3）面临互联网压力之下必须转型的传统企业需要与时俱进充分利用大数据的价值。

4. 大数据能做和不能做的事

1）大数据可以做到的事情

（1）诊断分析。

我们每天都在做这个事情。机器更擅长做这个。当一个事件发生的时候，我们发现对寻找起因感兴趣。比如，设想在沙漠 A 刮起了沙暴，我们有沙漠 A 地区的各种参数：温度、气压、骆驼、道路、汽车等等。如果我们能将这些参数跟该地区的沙暴联系起来，如果我们知道一些因果关系，可能就会避免沙暴。

（2）预测分析。

我们经常做这个事情。比如，我们在全球有一个酒店连锁。现在我们需要找出哪些酒店是没有达到销售目标的。如果知道相关信息，我们就可以将努力集中在那些目标身上。这成为预测分析的经典问题。

（3）在未知元素间寻找关联。

进行分析，在未知元素间寻找关联。比方说销售雇员的数量跟销售额真的没有关系吗？你可能会减少一些雇员来看看是否真的对销售额没有损失。

（4）规范的分析。

这是分析学的未来。比如说我们尝试着预测一个对大众目标的恐怖袭击然后安全地

将人们转移的策略，你需要做出在某个时候某个地点的游客人数以及可能会被爆炸所影响到的地区等各种预测。

（5）监控发生的事件。

行业中的大部分人都在做监控事件的工作。比如，你需要检测一个活动的反馈，找到强烈和不强烈的部分。这些分析将成为运营一个企业的关键。

2）大数据不可以做到的事情

（1）预测一个确定的未来。

使用机器学习的工具可以达到90%的精度，但是无法达到100%的准确。如果我们可以做到的话，我可以确切地告诉你谁才是目标以及每一次100%的响应率。但可惜的是这绝不会发生。

（2）归咎于新的数据源。

在任何分析上，数据处理耗费了大部分时间。我相信这就是你的创造力和商业理解的来源。但可能的是，你无法摆脱分析中最无聊的部分。

（3）找到一个商业问题的创新的解决方案。

创造力是人类永远的专利。没有机器可以找到问题的创新的解决方法。这是因为即使是人工智能也是由人们去编码的产物，创造力是不会从算法自己学习而来的。

（4）找到定义不是很明确的问题的解决方法。

分析学最大的挑战就是从业务问题中形成一个分析问题模型。如果你能做得很好，那么你正在成为一个分析明星。这种角色是机器无法取代的。比如，你的业务问题是管理损耗。除非定义了响应者、时间窗口等，没有预测算法可以帮你。

（5）数据管理/简化新数据源的数据。

随着数据量的增长，数据的管理正在成为一个难题。我们正在处理各种不同结构化的数据。比如，图表数据可能更适合网络分析，但是对活动数据是没用的。这部分信息也是机器无法分析的。

5. 大数据的分类

（1）按照数据分析的实时性，分为实时数据分析和离线数据分析两种。

① 实时数据分析。

实时数据分析一般用于金融、移动和互联网B2C等产品，往往要求在数秒内返回上亿行数据的分析，从而达到不影响用户体验的目的。要满足这样的需求，可以使用海量数据实时分析工具，采用精心设计的传统关系型数据库组成并行处理集群，或者采用一些内存计算平台，或者采用HDD的架构，这些无疑都需要比较高的软硬件成本。互联网企业的海量数据采集工具，均可以满足每秒数百“MB”的日志数据采集和传输需求，并将这些数据上载到中央系统上。

② 离线数据分析。

对于大多数反馈时间要求不是那么严苛的应用，比如离线统计分析、机器学习、搜索引擎的反向索引计算、推荐引擎的计算等，应采用离线分析的方式，通过数据采集工具将日志数据导入专用的分析平台。但面对海量数据，传统的数据处理工具往往会彻底失效，

主要原因是数据格式转换的开销太大,在性能上无法满足海量数据的采集需求。

(2) 按照大数据的数据量,分为内存级别、海量级别三种、商业智能(BI)级别。

① 内存级别。

这里的内存级别指的是数据量不超过集群的内存最大值。不要小看今天内存的容量,Facebook 缓存在内存的中的数据高达 320TB,而目前的 PC 服务器,内存也可以超过百“GB”。因此可以采用一些内存数据库,将热点数据常驻内存之中,从而取得非常快速的分析能力,非常适合实时分析业务。

② 海量级别。

海量级别指的是对于数据库和商业智能产品已经完全失效或者成本过高的数据量。海量数据级别的优秀企业级产品也有很多,但基于软硬件的成本原因,目前大多数互联网企业采用 Hadoop 的 HDFS 分布式文件系统来存储数据,并使用 MapReduce 进行分析。

③ 商业智能(BI)级别。

BI 级别指的是那些对于内存来说太大的数据量,但一般可以将其放入传统的 BI 产品和专门设计的 BI 数据库之中进行分析。目前主流的 BI 产品都有支持 TB 级以上的数据分析方案。

1.4　大数据技术的基本概念

1.4.1　传统数据处理

大数据处理数据时代理念的三大转变:要全体不要抽样,要效率不要绝对精确,要相关不要因果。具体的传统大数据处理方法其实有很多,但是根据长时间的实践,总结了一个基本的大数据处理流程,并且这个流程应该能够对大家理顺大数据的处理有所帮助。整个处理流程可以概括为四步,分别是采集、导入和预处理、统计和分析以及数据挖掘。

1. 采集

大数据的采集是指利用多个数据库来接收发自客户端的数据,并且用户可以通过这些数据库来进行简单的查询和处理工作。比如,电商会使用传统的关系型数据库 MySQL 和 Oracle 等来存储每一笔事务数据,除此之外,Redis 和 MongoDB 这样的 NoSQL 数据库也常用于数据的采集。

在大数据的采集过程中,其主要特点和挑战是并发数高,因为同时有可能会有成千上万的用户来进行访问和操作,比如火车票售票网站和淘宝,它们并发的访问量在峰值时达到上百万,所以需要在采集端部署大量数据库才能支撑。并且要对如何在这些数据库之间进行负载均衡和分片进行深入的思考和设计。

2. 统计/分析

统计与分析主要利用分布式数据库,或者分布式计算集群来对存储于其内的海量数据进行普通的分析和分类汇总等,以满足大多数常见的分析需求。在这方面,一些实时性需求会用到 Oracle 的 Exadata,以及基于 MySQL 的列式存储 Infobright 等,而一些批处

理，或者基于半结构化数据的需求可以使用 Hadoop。统计与分析这部分的主要特点和挑战是分析涉及的数据量大，对系统资源，特别是 I/O 会有极大的占用。

3. 导入/预处理

虽然采集端本身会有很多数据库，但是如果要对这些海量数据进行有效的分析，还是应该将这些来自前端的数据导入到一个集中的大型分布式数据库，或者分布式存储集群，并且可以在导入的基础上做一些简单的清洗和预处理工作。也有一些用户会在导入时使用来自推特(Twitter)的 Storm 来对数据进行流式计算，来满足部分业务的实时计算需求。导入与预处理过程的特点和挑战主要是导入的数据量大，每秒钟的导入量经常会达到百兆，甚至千兆级别。

4. 数据挖掘

与前面统计和分析过程不同的是，数据挖掘一般没有什么预先设定好的主题，主要是在现有数据上面进行基于各种算法的计算，起到预测的效果，从而实现一些高级别数据分析的需求。比较典型算法有用于聚类的 K-Means、用于统计学习的 SVM 和用于分类的 Naive Bayes，主要使用的工具有 Hadoop 的 Mahout 等。该过程的特点和挑战主要是用于挖掘的算法很复杂，并且计算涉及的数据量和计算量都很大，还有，常用数据挖掘算法都以单线程为主。

1.4.2 大数据分析的方法理论

越来越多的应用涉及大数据，这些大数据的属性，包括数量、速度、多样性等等都呈现了大数据不断增长的复杂性，所以，大数据的分析方法在大数据领域就显得尤为重要，可以说是决定最终信息是否有价值的决定性因素。基于此，大数据分析的方法理论有五个基本方面。

1. 预测性分析能力(Predictive Analytic Capabilities)

数据挖掘可以让分析员更好地理解数据，而预测性分析可以让分析员根据可视化分析和数据挖掘的结果做出一些预测性的判断。

2. 数据质量和数据管理(Data Quality and Data Management)

数据质量和数据管理是一些管理方面的最佳实践。通过标准化的流程和工具对数据进行处理，可以保证一个预先定义好的高质量的分析结果。

3. 可视化分析(Analytic Visualizations)

不管是对数据分析专家还是普通用户，数据可视化是数据分析工具最基本的要求。可视化可以直观地展示数据，让数据自己说话，让观众听到结果。

4. 语义引擎(Semantic Engines)

我们知道由于非结构化数据的多样性带来了数据分析的新的挑战，我们需要一系列的工具去解析、提取、分析数据。语义引擎需要被设计成能够从“文档”中智能提取信息。

5. 数据挖掘算法(Data Mining Algorithms)

可视化是给人看的,数据挖掘就是给机器看的。集群、分割、孤立点分析还有其他的算法让我们深入数据内部,挖掘有价值的信息。这些算法不仅要处理大数据的量,也要处理大数据的速度。

假如大数据真的是下一个重要的技术革新,那么我们最好把精力放在大数据能给我们带来的好处上,而不仅仅是挑战。

1.4.3　大数据技术

1. 大数据技术分类

大数据带来的不仅是机遇,同时也是挑战。传统的数据处理手段已经无法满足大数据的海量实时需求,需要采用新一代的信息技术来应对大数据的爆发。我们把大数据技术归纳为五大类,如表 1.1 所示。

表 1.1　大数据技术分类

大数据技术分类	大数据技术与工具
基础架构支持	云计算平台 云存储 虚拟化技术 网络技术 资源监控技术
数据采集	数据总线 ETL 工具
数据存储	分布式文件系统 关系型数据库 NoSQL 技术 关系型数据库与非关系型数据库融合 内存数据库
数据计算	数据查询、统计与分析 数据预测与挖掘 图谱处理 BI 商业智能
展现与交互	图形与报表 可视化工具 增强现实技术

1）基础架构支持

基础架构支持主要包括为支撑大数据处理的基础架构级数据中心管理、云计算平台、云存储设备及技术、网络技术、资源监控等技术。大数据处理需要拥有大规模物理资源的云数据中心和具备高效的调度管理功能的云计算平台的支撑。

2）数据采集技术

数据采集技术是数据处理的必备条件,首先需要有数据采集的手段,把信息收集上

来，才能应用上层的数据处理技术。数据采集除了各类传感设备等硬件软件设施之外，主要涉及的是数据的ETL（采集、转换、加载）过程，能对数据进行清洗、过滤、校验、转换等各种预处理，将有效的数据转换成适合的格式和类型。同时，为了支持多源异构的数据采集和存储访问，还需设计企业的数据总线，方便企业各个应用和服务之间数据的交换和共享。

3）数据存储技术

数据经过采集和转换之后，需要存储归档。针对海量的大数据，一般可以采用分布式文件系统和分布式数据库的存储方式，把数据分布到多个存储结点上，同时还需提供备份、安全、访问接口及协议等机制。

4）数据计算

我们把与数据查询、统计、分析、预测、挖掘、图谱处理、BI商业智能等各项相关的技术统称为数据计算技术。数据计算技术涵盖数据处理的方方面面，也是大数据技术的核心。

5）数据展现与交互

数据展现与交互在大数据技术中也至关重要，因为数据最终需要为人们所使用，为生产、运营、规划提供决策支持。选择恰当的、生动直观的展示方式能够帮助我们更好地理解数据及其内涵和关联关系，也能够更有效地解释和运用数据，发挥其价值。在展现方式上，除了传统的报表、图形之外，我们还可以结合现代化的可视化工具及人机交互手段，甚至是基于最新的如Google眼镜等增强现实手段，来实现数据与现实的无缝接口。

2. 三大技术推动大数据分析平台的发展

在互联网技术横行的时代，数据即价值，数据即资源。大数据分析工具的职责就是规整数据，挖掘价值。因此，大数据分析平台的发展在一定程度上代表着大数据的发展。而在现阶段，云存储技术、感知技术、数据可视化技术成为大数据应用技术中不可或缺的组成部分。

1）云存储技术

大数据可以抽象地分为大数据存储和大数据分析，这两者的关系是：大数据存储的目的是支撑大数据分析。大数据存储致力于研发可以扩展至PB甚至EB级别的大数据分析平台；大数据分析关注在最短的时间内处理大量不同类型的数据集。

根据著名的“摩尔定律”，18个月集成电路的复杂性就增加一倍。所以，存储器的成本大约每18～24个月就下降一半。这意味着云存储技术的潜力巨大，同时对于大数据分析平台而言，意味着更大的数据存储量和功能更强的线上大数据分析平台。

2）数据抓取技术

现在大多数的大数据分析平台的数据抓取功能还停留在对固定数据库的数据处理和整合上。但是随着互联网技术的应用拓展，直接从互联网甚至是行为个体上直接抓取数据并非是不可能的，在技术上也是可行的。

大数据的采集和数据抓取技术的发展是紧密联系的。以传感器技术、指纹识别技术、射频识别RFID技术、坐标定位技术等为基础的感知能力提升同样是物联网发展的基石。

而随着智能手机的普及,感知技术迎来了发展的高峰期。大数据分析平台未来极有可能整合数据抓取技术,变被动分析为主动寻找,从而攀上大数据分析技术发展的新高峰。

3)数据可视化技术

数据可视化技术是当下最热门的大数据应用数据,除了末端展示的需要,数据可视化也是数据分析时不可或缺的一部分,即返回数据时的二次分析。而数据可视化也利于大数据分析平台的学习功能建设,让没有技术背景和初学者也能很快掌握大数据分析平台的操作。

未来的大数据分析平台的承载平台也不可能固定在某一类平台,但是无论哪一类平台,数据分析和分析结果的末端展示都离不开数据可视化技术。其实与其说数据可视化技术是大数据应用技术发展的需要,不如说数据可视化技术简化了数据分析技术,从而让更多人可以走近大数据,使用大数据。

在大数据应用技术发展的历程中,还有许多技术伴随左右,但都没有以上这三大技术重要,因为它们直接勾勒了大数据分析平台的未来甚至是人类的未来。

在大数据概念中,目前还没有哪项单一技术能够满足所有应用需求。这些大数据技术或针对数字营销数据进行优化,或分析社交网络数据,再或者主要用已知数据来预防未知的风险,其应用领域比较具有针对性。

3. 大数据平台的三个重要的技术部分

我们可以将一套完整的大数据平台拆分成几个不同的技术领域。从宏观上来看,大数据平台包含了三个重要的技术部分。

1)数据交易技术

这一部分技术所从事的工作,是对一些传统的关系型数据或者非结构化数据进行处理,这些数据包括 ERP 应用、数据仓库应用、在线交易处理(OLTP)等。

2)数据交互技术

数据交互是第二类组成部分,它也是成长最迅速的一类大数据技术。数据交互技术主要是对社交网络、物联网设备和传感器、地理定位、影像文件、互联网点击、电子邮件等应用产生的数据进行处理。

3)数据处理技术

最后是对数据的处理。在这一部分中,包含了技术架构、计算方式等内容。知名的 Hadoop 平台就是其中的一分子。

另一方面,从微观层面,我们可以对大数据平台再进行更加细致的剖析。

(1)数据存储。

数据存储是大数据平台的根本,也是所有大数据技术中产品种类最多的一个组成部分。没有了存储平台,数据也就没有了载体。在数据存储的组成中,包括了高性能的内核式分布存储系统、用户级的分布式存储以及业务级别的数据存储。这其中不乏 Hadoop HDFS 这样的知名产品。

(2)数据同步。

这一部分技术主要用于将基础架构产生的数据内容进行转换,以完成数据处理、系统

监控等方面的操作。

(3) 数据开发。

顾名思义,数据开发技术主要承担了搭建大数据平台上层建筑的任务。其中涵盖了用户认证、数据鉴权、工作流、数据管理等多方面的任务。Facebook 为了更好地应用大数据技术,特别开发了名为 Facebook Insights 的产品,将大数据平台中的单元和属性抽离出来,以更好地掌控数据资源。

(4) 数据计算。

这一部分毫无疑问是一个大数据平台最为重要的技术核心。其承担了对海量数据进行再加工、再处理的任务。一般来说,可以将其分为离线计算与实时计算两种模式。

离线计算一般适用于对时间属性不敏感的应用,相对而言,其技术开发和构建的成本较低。但是由于离线计算需要数据同步技术对数据进行采集,过大的数据量会使得采集过程失败,因此目前用于离线计算的数据量还不能太大。

相较于离线计算,实时计算处理速度更快,但是其成本很高。目前实时计算大都用于金融、互联网等行业。

(5) 数据挖掘。

数据挖掘并不是一个新的技术,目前其发展已经非常成熟。在大数据的概念下,数据挖掘被赋予了新的意义。其所处理的数据类别越来越广泛,同时为了迎接海量数据,数据挖掘工具的性能也在不断提升。

在当今这个飞速发展的数字时代,大数据已经成为我们生活中必不可少的一部分。展望未来,围绕大数据还将有一些新的技术和商业模式诞生。数据将成为如同服装、汽车、家电或者是食物一样的商品,成为人们选购的对象。同时,精通大数据相关技术的数据科学家,也会成为一个新兴的职业类型,在新时代中扮演重要的角色。

4. 云平台与云存储

大数据的强大后台是云计算。简单地说,云计算包括三个部分:基础设施服务(Infrastructure-as-a-Service,IaaS)、平台服务(Platform-as-a-Service,PaaS)和软件服务(Software-as-a-Service,SaaS)。

1) 基础设施服务(IaaS)

基础设施服务是最基础的,它是云的一个服务端,用户可以通过互联网从计算机基础设施获得服务。IaaS 的大多数用户是科技公司,他们通常有很强的 IT 专长,想要利用计算机强大的计算功能,但是又不想负责安装和维护。

2) 平台服务(PaaS)

这是一个以云计算为基础的软件研发平台服务,公司可以利用这个平台在已有软件的基础上进一步发展或研发软件。PaaS 环境能够和一些软件开发工具结合,例如 Java、.NET、Python 等,更方便用户进行编码以及在网络上共享其程序编码。目前 PaaS 在云计算的市场份额是三个部分中最小的,主要被一些公司用来外包其基础设施。

3) 软件服务(SaaS)

是目前云计算中利用最多并且发展最成熟的一部分,它利用互联网提供软件服务,而

不需要被下载到用户端或者存储在一个数据中心。很多数据处理和文本处理软件，例如 Word 等，开始逐渐转向一些云计算的软件服务，比如 Google Apps、Microsoft Office 365 等。

云计算的三个部分如图 1.5 所示。

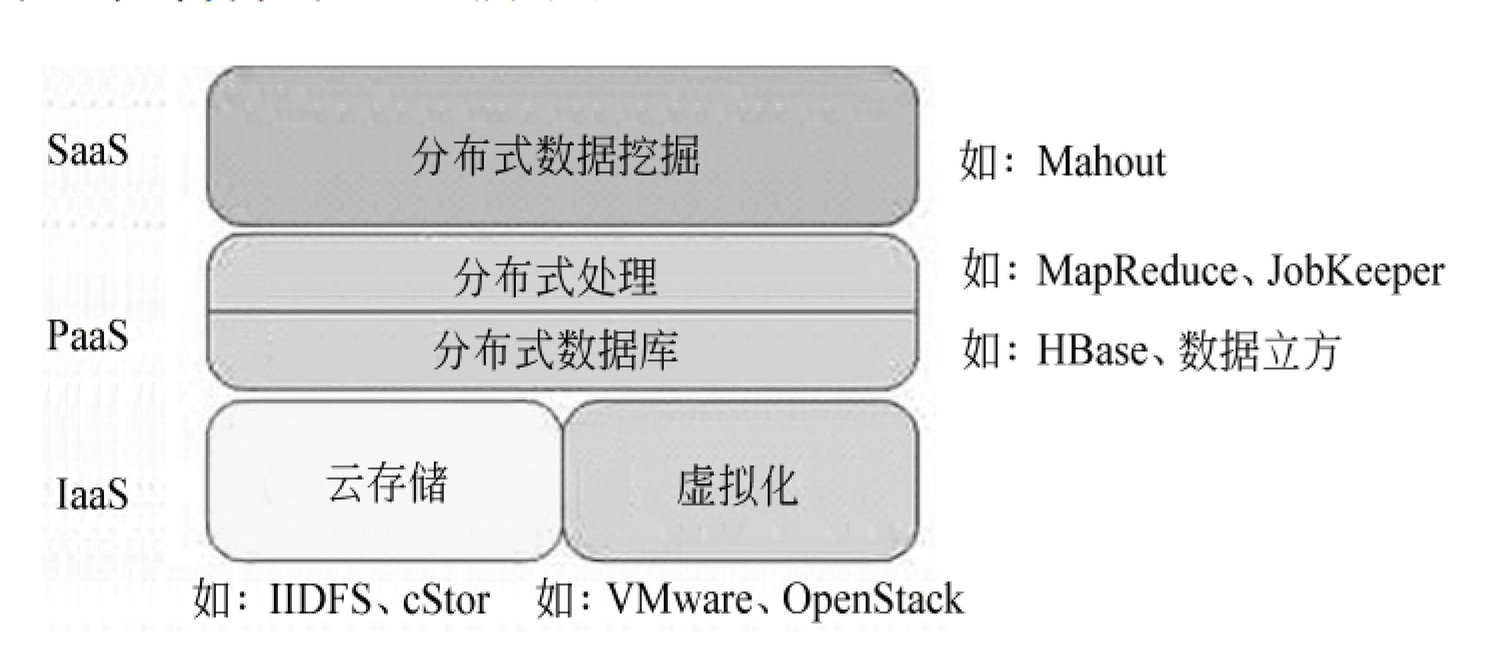


图 1.5　云计算的三个部分

云计算的三个部分有一些共同的特点。首先，用户不需要购买任何空间，而是采用租借的形式利用云端存储空间。第二，云计算服务提供商负责所有的维护、管理、空间计划、问题处理和后备存储等。最后，相比传统方法，云计算服务更方便、更快捷，IaaS 有更多的存储空间，PaaS 可以处理更多的平台服务，SaaS 可以被更多用户利用。

1.5　大数据的社会价值

1.5.1　大数据的社会价值体现

大数据技术的出现实现了巨大的社会价值，主要表现在如下几个方面。

1. 能够推动实现巨大经济效益

大数据技术的出现能够推动社会实现巨大经济效益，比如对中国零售业净利润增长的贡献，降低制造业产品开发、组装成本等。在 2013 年全球大数据直接和间接拉动信息技术支出达 1200 亿美元。

2. 能够推动增强社会管理水平

大数据在公共服务领域的应用，可有效推动相关工作开展，提高相关部门的决策水平、服务效率和社会管理水平，产生巨大的社会价值。欧洲多个城市通过分析实时采集的交通流量数据，指导驾车出行者选择最佳路径，从而改善城市交通状况。

3. 如果没有高性能的分析工具，大数据的价值就得不到释放

对大数据应用必须保持清醒认识，既不能迷信其分析结果，也不能因为其不完全准确而否定其重要作用。

(1) 由于各种原因，所分析处理的数据对象中不可避免地会包括各种错误数据、无用数据，加之作为大数据技术核心的数据分析、人工智能等技术尚未完全成熟，所以对计算机完成的大数据分析处理的结果，无法要求其完全准确。例如，Google 通过分析亿万用

户搜索内容能够比专业机构更快地预测流感暴发，但由于微博上无用信息的干扰，这种预测也曾多次出现不准确的情况。

(2) 必须清楚定位的是，大数据作用与价值的重点在于能够引导和启发大数据应用者的创新思维，辅助决策。简单而言，若是处理一个问题，通常人能够想到一种方法，而大数据能够提供十种参考方法，哪怕其中只有三种可行，也将解决问题的思路拓展了三倍。

所以，客观认识和发挥大数据的作用，不夸大、不缩小，是准确认知和应用大数据的前提。

1.5.2 大数据在政府管理方面的应用

政府数据资源丰富，应用需求旺盛，政府既是大数据发展的推动者，也是大数据应用的受益者。这一年，政府应用大数据更好地响应社会和经济指标变化，解决城市管理、安全管控、行政监管中的实际问题，预测判断事态走势等。对政府管理而言，大数据的价值在于提高决策科学化与管理精细化的水平。表 1.2 为部分政府管理领域大数据应用案例。

表 1.2 政府管理领域大数据应用案例

	背景内容	数据来源	作用效果
公安打击网络售假	淘宝联手上海、福建、浙江、湖南等地公安机关，运用大数据查获网售假冒运动鞋案件，涉案总价值 2150 余万元	淘宝数据和公安数据	各地警方共破案 5 起，捣毁犯罪团伙 1 个、捣毁销售、仓储窝点 7 处，现场缴获各类假冒“耐克”运动鞋 300 余双
缓解停车问题	SpotHero 是一个手机应用，能够根据用户的位置和目的地及路况，实时跟踪停车位数量变化	入网城市的可用车库或停车位，以及相对应的价格、时间、区间数据	能够实时监控华盛顿、纽约、芝加哥、巴尔的摩、波士顿、密尔沃基和纽瓦克七个城市的停车位
证监会调查内幕交易	已调查内幕交易线索 375 起，立案 142 起，分别比以往同期增长了 21%、33%	交易数据、企业信息和历史内幕交易数据等	已将涉嫌利用“银润投资”“圆城黄金”“爱施德”“焦作万方”等 43 家上市公司的内幕信息，从事非法交易的 125 名个人和 3 家机构移交公安机关
税务数据分析应用	增强对税务风险的监管和控制；对即将出现的风险点进行提示	登记、申报、缴款、集中度状况、增值税全部销售收入等数据	实现了对 45 家定点联系企业，近 5 万户分支机构实施税源监控、纳税评估
山西省农业厅	建设山西省“畜牧兽医大数据系统平台”和“山西省省级畜牧兽医大数据中心”	农业厅数据、天气数据、畜牧兽医机构数据等	利用大数据增强全省重大动物疫病防控能力和畜产品质量安全监管能力

数据来源：赛迪智库整理，2015.3

通过案例可见，政府部门一方面掌握了大量的基础数据资源；另一方面，在城市管理、安全管控、行政监管等领域的应用需求旺盛。大数据带来的是从政务信息公开，到数据整合共享，它超越了传统行政思维模式，推动政府从“经验治理”转向“科学治理”。

1.5.3　大数据在公共服务领域的应用

大数据在公共服务中的交通、医疗、教育、预测服务等领域得到广泛应用。随着第三方服务机构的参与，公众需求被不断挖掘，应用场景逐步丰富。表 1.3 为部分公共服务领域大数据应用的案例。

表 1.3　公共服务领域大数据应用案例

	背景内容	数据来源	作用效果
英国 NHS 糖尿病管理	英国每年有高达 77 亿英镑用于处理糖尿病并发症。通过数据分析干预，大量的糖尿病所带来的并发症是可以避免的	通过移动终端收集患者的生活起居数据、生理变化数据、用药数据、饮食数据、运动数据和医生诊断数据	对收集到的信息进行糖尿病风险等级评估，根据评估情况为每个患者制定适宜的个性化的糖尿病干预治疗方案
医疗平台 Healthtap	很多人已经开始选择使用移动智能终端进行医疗咨询，医疗健康行业的移动互联网普及率已超过 10%	减肥，锻炼，睡眠，戒烟等患者上传的个人习惯数据和健康情况及病史；症状，病情，药物、检测诊疗数据；就诊时短信、视频数据等	根据患者信息，为其提供医生推荐、药物推荐等服务。减少用户就诊时间，提高医生和患者的匹配度
智能学习应用“优答”	新东方和腾讯宣布成立合资公司“微学明日”。开发智能学习应用 APP“优答”	用户在优答上的学习行为，分析用户的学习效率、知识掌握薄弱的环节等，积累了每个用户的英语学习数据	目前智能拍照扫题准确率达到 80% 以上，响应速度在 10 秒以内
智能学习平台	“行为评价和诱导”的智能学习平台可以实现全球几十万人同步学习，共享全球优质教育资源	大量单个个体学习行为数据	总结群体的行为数据呈现出的规律，从而对学习者的学习行为进行自动的提示、诱导和评价
百度高考作文押题	高考作文成为社会关注的焦点，高考作文题目预测是老师和考生的急切需求	大量作文范文、海量的作文相关搜索数据、年度风云搜索信息、新闻数据、社会热点等	成功押中全国 18 套作文考题中的 12 套，成功率达到 66.7%
民生银行大数据项目	6 月份上线“阿拉丁”大数据在线平台	包括柜员系统、实物黄金、ATM、手机银行等 100 多个业务系统源数据	降低运营成本，控制风险，提高产品精准营销能力
中信银行信用卡	结合实时、历史数据进行全局分析，风险管理部门每天评估客户的行为，并决定对客户的信用额度在同一天进行调整	用户信用数据、支付数据、消费行为数据、还款数据、用户画像及历史数据等	提供了统一的客户视图，更有针对的进行营销，缩短营销活动配置的时间
阿里信贷	面对中小企业和个人贷款难问题，以及商业银行风险管控需求，阿里开发了信用评估大数据应用	交易时间、价格、购买数量，买方和卖方的年龄、性别、地址、兴趣爱好等	通过掌握的企业交易数据，借助大数据技术自动分析判定是否给予企业贷款

续表

	背 景 内 容	数 据 来 源	作 用 效 果
基于受众跨屏的 RTB 广告	原有互联网模式存在难以跨屏、Cookie 生命周期短、受众行为信息缺失等问题	电信运营商对“管道”中的数据进行解析，整合用户的 Cookie、IMEI、计费代码等数据	电信运营商自身业务的精确营销正在从“被动”变为“主动”模式，在降低成本的同时精准完成了业务推送
Climate Corporation 意外天气保险	天气对农业生产的影响较大，面对天气和气候变化而产生的个性化农业保险需求日益强烈	250 万个地点的气候测量数据和大型气候模型的每日预测，结合 1500 亿例土壤观察数据，生成 10 万亿个模拟气候数据点	预测未来可能对农业生产造成破坏的各种天气，农民可以根据这种预测来选择相应的农业保险
旅游预测	百度推出旅游预测产品，提供景区客流量预测、游客人口属性分析、游客兴趣挖掘、舆情分析等服务	依托百度海量的用户搜索行为数据、微博数据、位置数据等多维度旅游行业相关数据	百度已经与九寨沟、四川旅游局、山东旅游局已达成合作意向
油价预测	Esurance 推出一款名为 Fuelcaster 的 App，专门帮助车主们预测近期油价	从全国各地加油站收集的数据计算而成，车主只需输入区号，就可获得所在区域的油价预测	帮助车主预测近期油价，提供购买建议；显示周边 10 个加油站的油价对比

数据来源：赛迪智库整理，2015.3

通过案例可知，政府或第三方机构可以通过对交通、医疗、教育、天气等领域的大数据实时分析，提高对危机事件和未来趋势的预判能力，为实现更好、更科学的危机响应和事前决策提供了技术基础。

1.6 大数据的商业应用

1.6.1 商业大数据的类型和价值挖掘方法

1. 商业大数据的类型

商业大数据的类型大致可分为三类：

(1) 传统企业数据(Traditionalenterprisedata)。

传统企业数据包括 CRM systems 的消费者数据，传统的 ERP 数据，库存数据以及账目数据等。

(2) 机器和传感器数据(Machine-generated/sensor data)。

机器和传感器数据包括呼叫记录(CallDetail Records)、智能仪表、工业设备传感器、物联网传感设备、设备日志(通常是 Digital exhaust)、交易数据等。

(3) 社交数据(Socialdata)。

社交数据包括用户行为记录、反馈数据等，如推特(Twitter)、脸书(Facebook)这样的社交媒体平台。

2. 大数据挖掘商业价值的方法

大数据挖掘商业价值的方法主要分为四种：

(1) 客户群体细分，为每个群体量定制特别的服务。

(2) 模拟现实环境，发掘新的需求同时提高投资的回报率。

(3) 加强部门联系，提高整条管理链条和产业链条的效率。

(4) 降低服务成本，发现隐藏线索进行产品和服务的创新。

3. 传统商业智能技术与大数据应用的比较

随着新型商业智能的产生，传统针对海量数据的存储处理，通过建立数据中心，建设包括大型数据仓库及其支撑运行的软硬件系统，设备(包括服务器、存储、网络设备等)越来越高档，数据仓库、OLAP 及 ETL、BI 等平台越来越庞大，但这些需要的投资越来越大，而面对数据的增长速度，越来越力不从心，所以基于传统技术的数据中心建设、运营和推广难度越来越大。

另外一般能够使用传统的数据库、数据仓库和 BI 工具能够完成的处理和分析挖掘的数据，还不能称为大数据，这些技术也不能叫大数据处理技术。面对大数据环境，包括数据挖掘在内的商业智能技术正在发生巨大的变化。

传统的传统商业智能技术，包括数据挖掘，主要任务是建立比较复杂的数据仓库模型、数据挖掘模型，来分析和处理不太多的数据。

由于云计算模式、分布式技术和云数据库技术的应用，我们不需要这么复杂的模型，不用考虑复杂的计算算法，就能够处理大数据，对于不断增长的业务数据，用户也可以通过添加低成本服务器甚至是 PC 也可以，来处理海量数据记录的扫描、统计、分析、预测。如果商业模式变化了，需要一分为二，那么新商业智能系统也可以很快地、相应地一分为二，继续强力支撑商业智能的需求。

所以实际是对传统商业智能的发展和促进，商业智能将出现新的发展机遇，面对风云变幻的市场环境，快速建模，快速部署是新商业智能平台的强力支撑。而不像过去那样艰难前行，难以承受商业运作的变化。大数据蕴含的商机如图 1.6 所示。

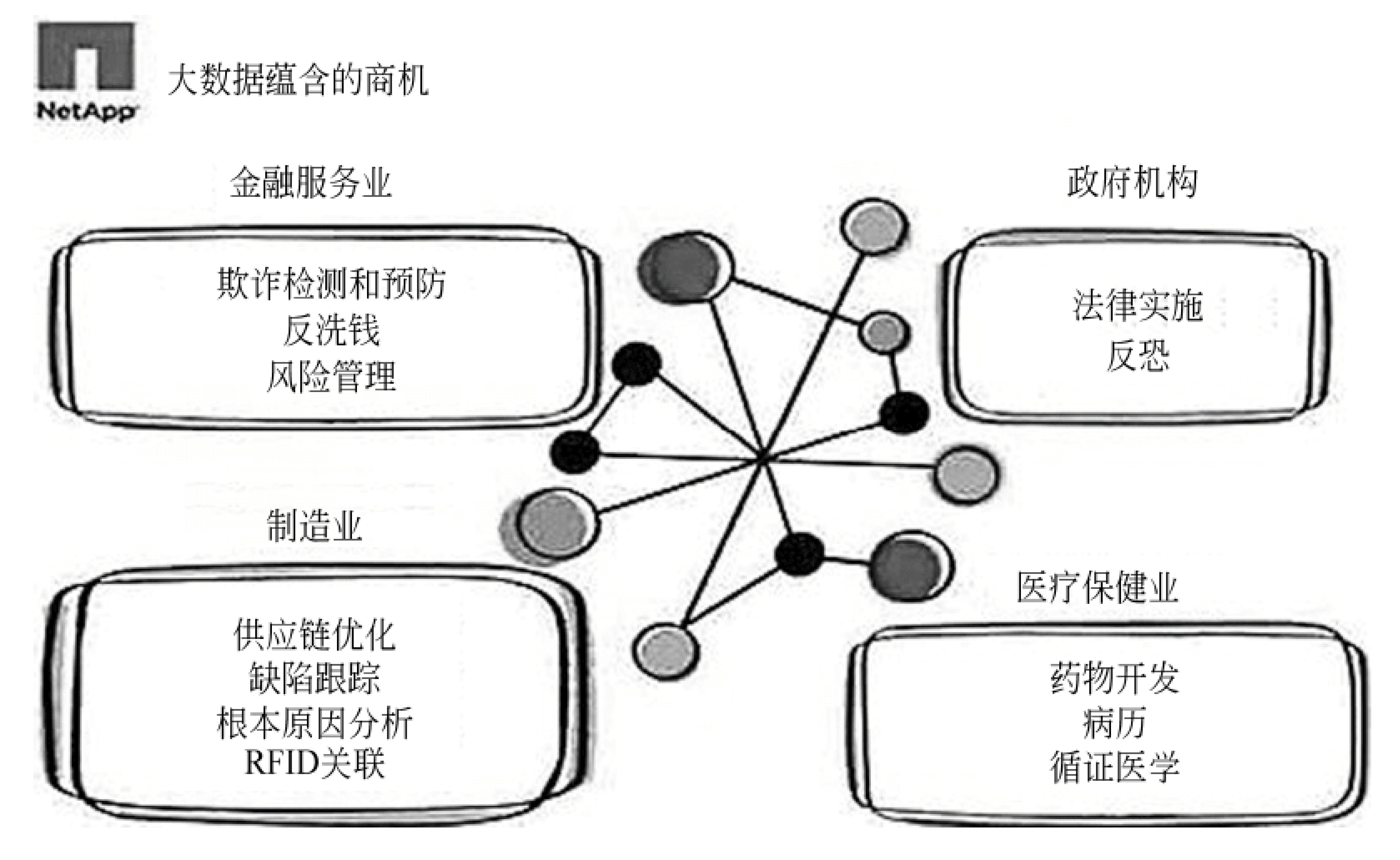


图 1.6　大数据蕴含的商机

1.6.2 全球大数据市场结构

全球大数据市场结构从垄断竞争向完全竞争格局演化。企业数量迅速增多,产品和服务的差异度增大,技术门槛逐步降低,市场竞争越发激烈。

全球大数据市场中,行业解决方案、计算分析服务、存储服务、数据库服务和大数据应用为市场份额排名最靠前的细分市场,分别占据35.4%、17.3%、14.7%、12.5%和7.9%的市场份额。云服务的市场份额为6.3%,基础软件占据3.8%的市场份额,网络服务仅占据了2%的市场份额。2011—2017年全球大数据细分领域市场规模及预测(单位:亿美元)见表1.4。

表1.4 2011—2017年全球大数据细分领域市场规模及预测 (单位:亿美元)

细分领域	2011年	2012年	2013年	2014年	2015年	2016年	2017年
云	3.6	6.2	11.9	18.2	25.2	30.5	36.5
行业解决方案	28	44.2	61.5	101	135	160	172
应用	5.2	9.9	16.9	34.5	52.9	66.5	77.5
非关系型数据库	0.7	1.3	2.9	5	8	10	12
关系型数据库	6.2	8.8	13.1	17.5	22.5	24.5	27
基础软件	1.4	4.4	8.3	10.8	12.5	16	19
网络	1.5	2.3	4.2	6.5	8.5	10.1	11.5
存储	11	17.5	30.9	42	55	64	69.5
计算	15.3	22.9	36.5	49.2	64	71	76

数据来源:Wikibon公司数据,2014.5

全球大数据发展呈现两极分化的态势。欧美等发达国家拥有先发优势,处于产业发展领导地位,中国、日本、韩国、澳大利亚、新加坡等国家分别发挥各自在数据资源、行业应用、技术积累、政策扶持等方面的优势,紧紧跟随,并在个别领域处于领先。其他多数国家的大数据发展相对缓慢,还停留在概念炒作和基础设施建设阶段。在开源技术的支撑下,技术已不是大数据发展的最大障碍,信息化基础和数据资源成为一个国家和地区大数据发展的关键要素。

1.6.3 中国大数据市场

我国大数据市场的供给结构初步形成,并与全球市场相似,呈现三角形结构,即以百度、阿里、腾讯为代表的互联网企业,以华为、联想、浪潮、曙光、用友等为代表的传统IT厂商,以亿赞普、拓尔思、海量数据、九次方等为代表的大数据企业。我国大数据市场的供给结构如图1.7所示。

国内外大数据产业链重要企业列表如表1.5所示。

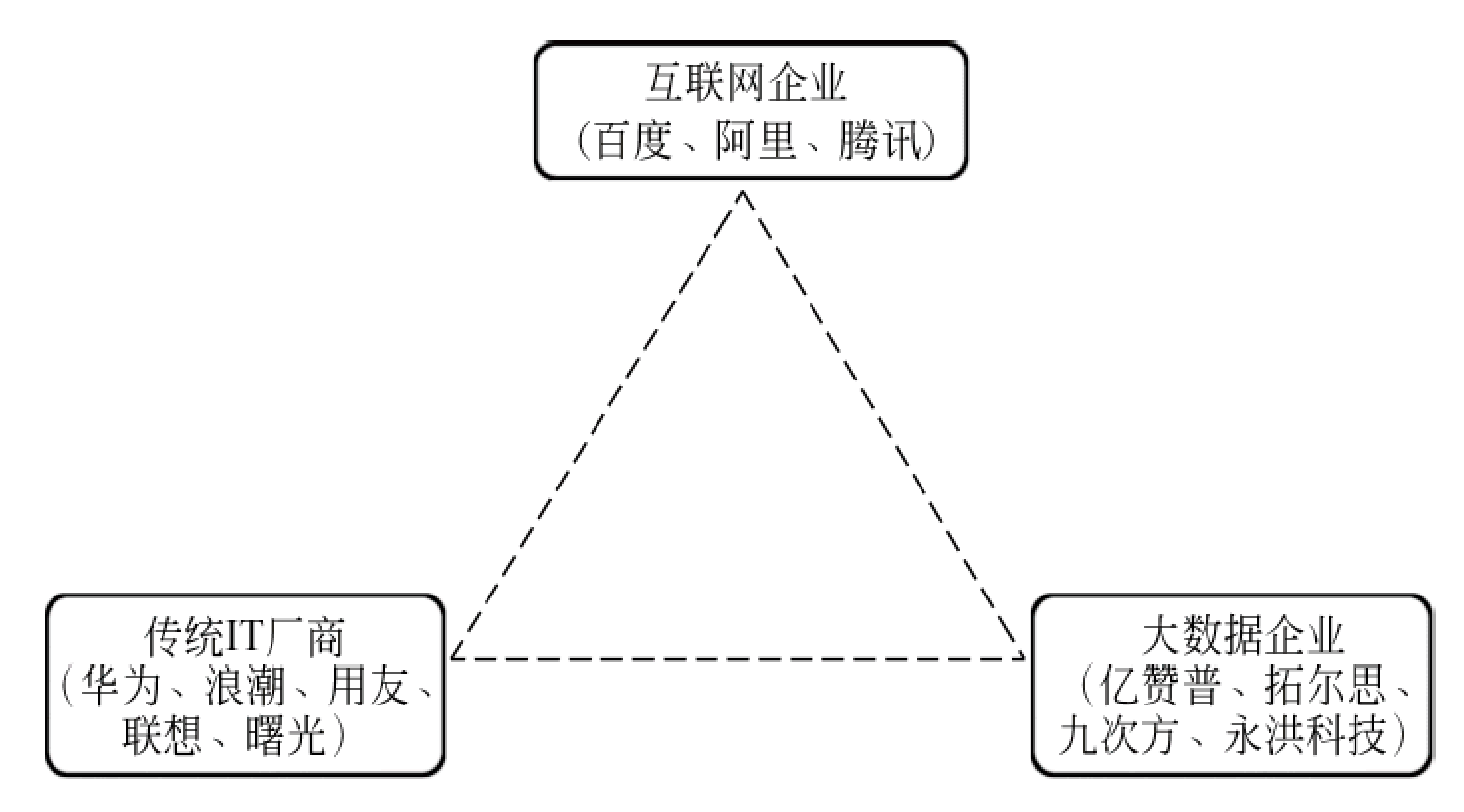


图 1.7　我国大数据市场供给结构图

表 1.5　大数据产业链重要企业名录

产业环节	国外代表企业	国内代表企业
大数据处理平台	IBM、微软、MapR、Zettaset、Cloudera、HStreaming、Hadapt、DataStax、Datameer	百度、阿里、腾讯
数据获取	DataSift、Gnip、Knoema、Infochimps、SpaceCurve、Windows Azure Marketplace	华胜天成、用友软件
数据存储	10gen、DataStax、CouchBase、Neo4j、Cloudant、Marklogic、HP Vertica、IBM、Netezza、Teradata	忆恒创源、永洪科技、百度、华胜天成、拓尔思、东方国信、博彦科技
数据处理和分析	Palantir、Platfora、Pervasive、Datameer、MetaMarkets	天地超云、联想、永洪科技、东方国信、百度、天源迪科、亿赞普
数据应用	Rocketfuel、Tapad、Yieldbot、Chartbeat、Lattice engines、谷歌、亚马逊、Ravel、23andMe	百度、阿里、腾讯、云端时代、华胜天成、灵玖软件、天云融创、品友互动
数据安全	DataGuise、Stormpath、Imperva、Dataguise	蓝盾、启明星辰、奇虎

1.6.4　大数据给中国带来的十大商业应用场景

在未来的几十年里，大数据都将会是一个重要的话题。大数据影响着每一个人，并在可以预见的未来继续影响着。大数据冲击着许多主要行业，包括零售业、金融行业、医疗行业等，大数据也将彻底地改变我们的生活。下面就来看看大数据给中国带来的十大商业应用场景，未来大数据产业将会是一个万亿市场。

1. 智慧城市

如今，世界超过一半的人口生活在城市里，到 2050 年这一数字会增长到 75%。政府需要利用一些技术手段来管理好城市，使城市里的资源得到良好配置。既不出现由于资源配置不平衡而导致的效率低下，又要避免不必要的资源浪费而导致的财政支出过大。大数据作为其中的一项技术可以有效帮助政府实现资源科学配置，精细化运营城市，打造智慧城市。

城市的道路交通，完全可以利用GPS数据和摄像头数据来进行规划，包括道路红绿灯时间间隔和关联控制，包括直行和左右转弯车道的规划、单行道的设置。利用大数据技术实施的城市交通智能规划，至少能够提高30%左右的道路运输能力，并能够降低交通事故率。在美国，政府依据某一路段的交通事故信息来增设信号灯，降低了50%以上的交通事故率。机场的航班起降依靠大数据将会提高航班管理的效率，航空公司利用大数据可以提高上座率，降低运行成本。铁路利用大数据可以有效安排客运和货运列车，提高效率、降低成本。

城市公共交通规划、教育资源配置、医疗资源配置、商业中心建设、房地产规划、产业规划、城市建设等都可以借助于大数据技术进行良好规划和动态调整。

大数据技术可以了解经济发展情况，各产业发展情况，消费支出和产品销售情况，依据分析结果，科学地制定宏观政策，平衡各产业发展，避免产能过剩，有效利用自然资源和社会资源，提高社会生产效率。大数据技术也能帮助政府进行支出管理，透明合理的财政支出将有利于提高公信力和监督财政支出。大数据及大数据技术带给政府的不仅仅是效率提升、科学决策、精细管理，更重要的是数据治国、科学管理的意识改变，未来大数据将会从各个方面来帮助政府实施高效和精细化管理，具有极大的想象空间。

2. 金融行业

大数据在金融行业应用范围较广，典型的案例有花旗银行利用IBM电脑为财富管理客户推荐产品，美国银行利用客户点击数据集为客户提供特色服务。中国金融行业大数据应用开展的较早，但都是以解决大数据效率问题为主，很多金融行业建立了大数据平台，对金融行业的交易数据进行采集和处理。

金融行业过去的大数据应用以分析自身财务数据为主，以提供动态财务报表为主，以风险管理为主。在大数据价值变现方面，开展得不够深入，这同金融行业每年上万亿的净利润相比是不匹配的。现在已经有一些银行和证券开始和移动互联网公司合作，一起进行大数据价值变现，其中招商银行、平安集团、兴业银行、国信证券、海通证券在移动大数据精准营销、获客、用户体验等方面进行了不少的尝试，大数据价值变现效果还不错，大数据正在帮助金融行业进行价值变现。大数据在金融行业的应用可以总结为以下五个方面。

(1) 精准营销：依据客户消费习惯、地理位置、消费时间进行推荐。

(2) 风险管控：依据客户消费和现金流提供信用评级或融资支持，利用客户社交行为记录实施信用卡反欺诈。

(3) 决策支持：利用决策树技术进抵押贷款管理，利用数据分析报告实施产业信贷风险控制。

(4) 效率提升：利用金融行业全局数据了解业务运营薄弱点，利用大数据技术加快内部数据处理速度。

(5) 产品设计：利用大数据计算技术为财富客户推荐产品，利用客户行为数据设计满足客户需求的金融产品。

3. 医疗行业

医疗行业拥有大量病例、病理报告、医疗方案、药物报告等。如果这些数据进行整理

和分析，将会极大地帮助医生和病人。在未来，借助于大数据平台我们可以收集疾病的基本特征、病例和治疗方案，建立针对疾病的数据库，帮助医生进行疾病诊断。

如果未来基因技术发展成熟，可以根据病人的基因序列特点进行分类，建立医疗行业的病人分类数据库。在医生诊断病人时可以参考病人的疾病特征、化验报告和检测报告，参考疾病数据库来快速帮助病人确诊。在制定治疗方案时，医生可以依据病人的基因特点，调取相似基因、年龄、人种、身体情况相同的有效治疗方案，制定出适合病人的治疗方案，帮助更多人及时进行治疗。同时这些数据也有利于医药行业开发出更加有效的药物和医疗器械。

医疗行业的数据应用一直在进行，但是数据没有打通，都是孤岛数据，没有办法大规模应用。未来需要将这些数据统一收集起来，纳入统一的大数据平台，为人类健康造福。政府是推动这一趋势的重要动力，未来市场将会超过几千亿元。

4. 农牧业

农产品不容易保存，合理种植和养殖农产品对农民非常重要。借助于大数据提供的消费能力和趋势报告，政府将为农牧业生产进行合理引导，依据需求进行生产，避免产能过剩，造成不必要的资源和社会财富浪费。大数据技术可以帮助政府实现农业的精细化管理，实现科学决策。在数据驱动下，结合无人机技术，农民可以采集农产品生长信息、病虫害信息。

农业生产面临的危险因素很多，但这些危险因素很大程度上可以通过除草剂、杀菌剂、杀虫剂等技术产品进行消除。天气成了影响农业非常大的决定因素。过去的天气预报仅仅能提供当地的降雨量，但农民更关心有多少水分可以留在土地上，这些是受降雨量和土质来决定的。Climate 公司利用政府开放的气象站的数据和土地数据建立了模型，可以告诉农民可以在哪些土地上耕种，哪些土地今天需要喷雾并完成耕种，哪些正处于生长期的土地需要施肥，哪些土地需要 5 天后才可以耕种，大数据技术可以帮助农业创造巨大的商业价值。

5. 零售行业

零售行业比较有名气的大数据案例就是沃尔玛的啤酒和尿布的故事，以及 Target 通过向年轻女孩寄送尿布广告而告知其父亲女孩怀孕的故事。

零售行业可以通过客户购买记录，了解客户关联产品购买喜好，将相关的产品放到一起增加来增加产品销售额，例如将洗衣服相关的化工产品例如洗衣粉、消毒液、衣领净等放到一起进行销售。根据客户相关产品购买记录而重新摆放的货物将会给零售企业增加 30% 以上的产品销售额。

零售行业还可以记录客户购买习惯，将一些日常需要的必备生活用品，在客户即将用完之前，通过精准广告的方式提醒客户进行购买。或者定期通过网上商城进行送货，既帮助客户解决了问题，又提高了客户体验。

电商行业的巨头——天猫和京东，已经通过客户的购买习惯，将客户日常需要的商品例如尿不湿、卫生纸、衣服等商品依据客户购买习惯事先进行准备。当客户刚刚下单，商

品就会在24小时内或者30分钟内送到客户门口，提高了客户体验，让客户连后悔的时间都没有。

利用大数据的技术，零售行业将至少会提高30%左右的销售额，并提升客户购买体验。

6. 大数据技术产业

进入移动互联网之后，非结构化数据和结构化数据呈指数方式增长。现在人类社会每两年产生的数据将超过人类历史过去所有数据之和。这些数据如何存储和处理将会成为很大的问题。

这些大数据为大数据技术产业提供了巨大的商业机会。据估计全世界在大数据采集、存储、处理、清晰、分析所产生的商业机会将会超过2000亿美元，包括政府和企业在大数据计算和存储，数据挖掘和处理等方面等投资。中国2014年大数据产业产值已经超过了千亿人民币，贵阳大数据博览会就吸引了400多家厂商来参展，充分说明大数据产业的未来的商业价值巨大。

未来中国的大数据产业将会呈几何级数增长，在5年之内，中国的大数据产业将会形成万亿规模的市场。不仅仅是大数据技术产品的市场，也将是大数据商业价值变现的市场。大数据将会在企业的精准营销、决策分析、风险管理、产品设计、运营优化等领域发挥重大的作用。

大数据技术产业将会解决大数据存储和处理的问题，大数据服务公司将利用自身的数据将解决大数据价值变现问题，其所带来的市场规模将会超过千亿人民币。中国目前拥有大数据，并提供大数据价值变现服务的公司除了众所周知的BAT和移动运营商之外，360、小米、京东等都会成为大数据价值变现市场的有力参与者，期望他们将市场进一步做大，帮助所有企业实现大数据价值变现。

7. 物流行业

中国的物流产业规模大概有5万亿元左右，其中公里物流市场大概有3万亿元左右。物流行业的整体净利润从过去的30%以上降低到了20%左右，并且下降的趋势明显。物流行业很多的运力浪费在返程空载、重复运输、小规模运输等方面。中国市场最大等物流公司所占的市场份额不到1%。因此资源需要整合，运送效率需要提高。

物流行业借助于大数据，可以建立全国物流网络，了解各个结点的运货需求和运力，合理配置资源，降低货车的返程空载率，降低超载率，减少重复路线运输，降低小规模运输比例。通过大数据技术，及时了解各个路线货物运送需求，同时建立基于地理位置和产业链的物流港口，实现货物和运力的实时配比，提高物流行业的运输效率。借助于大数据技术对物流行业进行的优化资源配置，至少可以增加物流行业10%左右的收入，其市场价值将在5000亿元左右。

8. 房地产业

中国房地产业发展的高峰已经过去，其面临的挑战逐渐增加，房地产业正从过去的粗

放发展方式转向精细运营方式，房地产企业在拍卖土地、住房地产开发规划、商业地产规划方面也将会谨慎进行。

借助于大数据，特别是移动大数据技术。房地产业可以了解开发土地所在范围常住人口数量、流动人口数量、消费能力、消费特点、年龄阶段、人口特征等重要信息。这些信息将会帮助房地产商在商业地产开发、商户招商、房屋类型、小区规模进行科学规划。利用大数据技术，房地产行业将会降低房地产开发前的规划风险，合理制定房价，合理制定开发规模，合理进行商业规划。大数据技术可以降低土地价格过高、实际购房需求过低的风险。已经有房地产公司将大数据技术应用于用户画像、土地规划、商业地产开发等领域，并取得了良好的效果。

9. 制造业

制造业过去面临生产过剩的压力，很多产品包括家电、纺织产品、钢材、水泥、电解铝等都没有按照市场实际需要生产，造成了资源的极大浪费。利用电商数据、移动互联网数据、零售数据，我们可以了解未来产品市场都需求，合理规划产品生产，避免生产过剩。

例如，依据用户在电商搜索产品的数据以及物流数据，可以推测出家电产品和纺织产品未来的实际需求量，厂家将依据这些数据来进行生产，避免生产过剩。移动互联网的位置信息可以帮助了解当地人口进出的趋势，避免生产过多的钢材和水泥。

大数据技术还可以根据社交数据和购买数据来了解客户需求，帮助厂商进行产品开发，设计和生产出满足客户需要的产品。

10. 互联网广告业

2014 年中国互联网广告市场迎来发展高峰，市场规模预计达到 1500 亿元左右，较 2013 年增长 56.5%。数字广告越来越受到广告主的重视，其未来市场规模越来越大。2014 年美国的互联网广告市场规模接近 500 亿美元，参考中国的人口消费能力，其市场规模会很快达到 2000 亿元人民币左右。

过去到广告投放都是以好的广告渠道＋广播式投放为主，广告主将广告交给广告公司，由广告公司安排投放，其中 SEM 广告市场最大，其他的广告投放方式也是以页面展示为主，大多是广播式广告投放。广播式投放的弊端是投入资金大，没有针对目标客户，面对所有客户进行展示，广告的转化率较低，并存在数字广告营销陷阱等问题。

大数据技术可以将客户在互联网上的行为记录下来，对客户的行为进行分析，打上标签并进行用户画像。特别是进入移动互联网时代之后，客户主要的访问方式转向了智能手机和平板电脑，移动互联网的数据包含了个人的位置信息，其 360 度用户画像更加接近真实人群。360 度用户画像可以帮助广告主进行精准营销，广告公司可以依据用户画像的信息，将广告直接投放到用户的移动设备，通过用户经常使用的 APP 进行广告投放，其广告的转化可以大幅度提高。利用移动互联网大数据技术进行的精准营销将会提高十倍以上的客户转化率，广告行业的程序化购买正在逐步替代广播式广告投放。大数据技术将帮助广告主和广告公司直接将广告投放给目标用户，从而降低广告投入，提高广告的转化率。

1.7 大数据与商业模式创新

1.7.1 商业模式的创新特点

商业模式创新的企业有几个共同特征，或者说构成商业模式创新的特点。

(1) 商业模式创新更注重从客户的角度，从根本上思考设计企业的行为，视角更为外向和开放，更多注重和涉及企业经济方面的因素。

商业模式创新的出发点，是如何从根本上为客户创造增加的价值。因此，其逻辑思考的起点是客户的需求，根据客户需求考虑如何有效满足它，这点明显不同于许多技术创新。用一种技术可能有多种用途，技术创新的视角常常是从技术特性与功能出发，看它能用来干什么，去找它潜在的市场用途。商业模式创新即使涉及技术，也多是和技术的经济方面因素，与技术所蕴含的经济价值及经济可行性有关，而不是纯粹的技术特性。

(2) 商业模式创新表现得更为系统和根本，它不是单一因素的变化。它常常涉及商业模式多个要素同时大的变化，需要企业组织的较大战略调整，是一种集成创新。商业模式创新往往伴随产品、工艺或者组织的创新；反之，则未必足以构成商业模式创新。

如开发出新产品或者新的生产工艺，就是通常认为的技术创新。技术创新，通常是对有形实物产品的生产来说的。但如今是服务为主导的时代，如美国2006年服务业比重高达68.1%，对传统制造企业来说，服务也远比以前重要。因此，商业模式创新也常体现为服务创新，表现为服务内容及方式及组织形态等多方面的创新变化。

(3) 从绩效表现看，商业模式创新如果提供全新的产品或服务，那么它可能开创了一个全新的可赢利产业领域，即便提供已有的产品或服务，也更能给企业带来更持久的赢利能力与更大的竞争优势。

传统的创新形态，能带来企业局部内部效率的提高和成本的降低，而且它容易被其他企业在较短期时期模仿。商业模式创新，虽然也表现为企业效率提高、成本降低，由于它更为系统和根本，涉及多个要素的同时变化，因此，它也更难以被竞争者模仿，常给企业带来战略性的竞争优势，而且优势常可以持续数年。

1.7.2 商业模式创新可以为企业带来什么

1. 战略定位创新

战略定位创新主要是围绕企业的价值主张、目标客户及顾客关系方面的创新，具体指企业选择什么样的顾客，为顾客提供什么样的产品或服务，希望与顾客建立什么样的关系，其产品和服务能向顾客提供什么样的价值等方面的创新。在激烈的市场竞争中，没有哪一种产品或服务能够满足所有的消费者，战略定位创新可以帮助我们发现有效的市场机会，提高企业的竞争力。在战略定位创新中，企业首先要明白自己的目标客户是谁，其次是如何让企业提供的产品或服务在更大程度上满足目标客户的需求，在前两者都确定的基础上，再分析选择何种客户关系。合适的客户关系也可以使企业的价值主张更好地满足目标客户。

2. 资源能力创新

资源能力创新是指企业对其所拥有的资源进行整合和运用能力的创新，主要是围绕企业的关键活动，建立和运转商业模式所需要的关键资源的开发和配置、成本及收入源方面的创新。所谓关键活动，是指影响其核心竞争力的企业行为；关键资源指能够让企业创造并提供价值的资源，主要指那些其他企业不能够代替的物质资产、无形资产、人力资本等。在确定了企业的目标客户、价值主张及顾客关系之后，企业可以进一步进行资源能力的创新。

战略定位是企业进行资源能力创新的基础，而且资源能力创新的四个方面也是相互影响的。一方面，企业要分析在价值链条上自己拥有或希望拥有那些别人不能代替的关键能力，根据这些能力进行资源的开发与配置；另一方面，如果企业拥有某项关键资源如专利权，也可以针对其关键资源制定相关的活动；对关键能力和关键资源的创新也必将引起收入源及成本的变化。

3. 商业生态环境创新

商业生态环境创新是指企业将其周围的环境看作一个整体，打造出一个可持续发展的共赢的商业环境。商业生态环境创新主要围绕企业的合作伙伴进行创新，包括供应商、经销商及其他市场中介，在必要的情况下，还包括其竞争对手。市场是千变万化的，顾客的需求也在不断变化，单个企业无法完全完成这一任务，企业需要联盟、需要合作来达到共赢。

企业战略定位及内部资源能力都是企业建立商业生态环境的基础。没有良好的战略定位及内部资源能力，企业将失去挑选优秀外部合作者的机会以及与他们议价的筹码。一个可持续发展的共赢的商业环境也将为企业未来发展及运营能力提供保证。

4. 混合商业模式创新

混合商业模式创新是一种战略定位创新、资源能力创新和商业生态环境创新相互结合的方式。据研究，企业的商业模式创新一般都是混合式的，因为企业商业模式的构成要素战略定位、内部资源、外部资源环境之间是相互依赖、相互作用的，每一部分的创新都会引起另一部分相应的变化。而且，这种由战略定位创新、资源能力创新和商业能力创新两两相结合甚至同时进行的创新方式，都会为企业经营业绩带来巨大的改善。

1.7.3　基于大数据分析的商业模式创新

1. 加大数据处理分析能力

所谓大数据，最为核心的就要看对于大量数据的核心分析能力。但是，大数据核心分析能力的影响不仅存在于数据管理策略、数据可视化与分析能力等方面，从根本上也对数据中心 IT 基础设施架构甚至机房设计原则等提出了更高的要求。为了达到快速高效的处理大量数据的能力，整个 IT 基础设施需要进行整体优化设计，应充分考量后台数据中心的高节能性、高稳定性、高安全性、高可扩展性、高度冗余、基础设施建设这六个方面，同时更需要解决大规模结点数的数据中心的部署、高速内部网络的构建、机房散热以及强大

的数据备份等问题。

2. 提高专业技术人员的技术水平

有这样一则故事，讲的是福特爱“才”、取之有道的故事：有一次福特公司的一台马达坏了，公司出动所有的工程技术人员，但是没有一个人能修复，福特公司只得另请高明。几经寻找，找到了坦因曼思，他原是德国工程技术人员，流落到美国后，被一家小工厂的老板看中并雇用了他。

他到了现场后，在马达旁听了听，要了把梯子，一会儿爬上一会爬下，最后在马达的一个部位用粉笔画一道线，写上几个字“这儿的线圈多了 16 圈”。果然把多余的线圈去掉，马达立即恢复正常。亨利·福特非常赏识坦因曼思的才华，就邀请他来福特公司工作，但坦因曼思却说：“我现在的公司对我很好，我不能忘恩负义”。福特马上说：“我把你供职的公司买下来，你就可以来工作了。”福特为了得到一个人才不惜买下了一个公司。

由此可见人才的重要性，因此企业要采取多种形式引进优秀人才。在注重优秀人才引进的同时加强对人才的教育和培养。建立合理的人力资源管理体制。建立起合理的薪酬制度和员工激励制度。中小企业可以积极满足员工的各种需要，促进组织目标实现的福利项目。比如医疗福利等，为员工提供一个自我发展的舞台、自我价值实现的桥梁。

同时，还可以借鉴在西方国家盛行的“弹性福利计划”，由员工在企业规定的时间和金额范围内，按照自己的意愿搭建自己的福利项目组合，满足员工对福利灵活机动的要求，提高员工的满意度，最终实现留住优秀人才的长远发展目标。

3. 理论与实践相结合促进商业模式的创新

阿里巴巴是全球企业界电子商务的著名品牌，是目前全球最大的网上交易市场和商务交流社区。良好的定位、稳固的结构、优秀的服务使阿里巴巴为全球首家拥有 600 余万商人的电子商务网站，成为全球商人网络推广的首选网站，被商人们评委“最受欢迎的 B2B 网站”。阿里巴巴商业模式创新的成功主要可归功于其相对完善的网上诚信保障机制的建立。

(1) 精准的市场定位。

阿里巴巴清晰地为业界定位其目标客户——众多的中小企业。阿里巴巴相关人士认为：在全球化日益发展的今天，中小企业无疑将拥有更多的介入机会和发展动力，依靠自身激动灵活的优势获得更大的成长空间。

(2) 关键资源能力的构建。

一是团队智慧。阿里巴巴团队认为，帮助客户合同是成功，才是自己成功的最好体现。二是文化资源。阿里巴巴共享价值观体系的强大企业文化可归纳为六个核心价值观，即客户第一、团队合作、拥抱变化、诚信、激情、敬业。

(3) 成功的盈利模式。

阿里巴巴的利润主要来源于注册会员缴纳的会员费。其付费会员有两种类型：国际交易平台的会员和国内交易平台的会员。

1.8　如何成为“大数据企业”

对企业而言，大数据实质上是一种管理思维，其支点在于业务信息资源与社交媒体的融合，以及内外部数据的融合，在这样的支点上反思企业的组织形态、运作范式和价值创造模式，是“大数据企业”的真正内涵所在。

一家中等规模的百货商场，通过视频监控记录商场各个区域的客流人数，从而评估每天各个时段客流的在店时长，进而结合销售记录数据估算出客流中带有明确购买目标的“搜索型”顾客和无明确购买目标的“浏览型”顾客的比例，从而为之设计有针对性的营销手段和服务措施。

这一实践中所涉及的数据量，从技术视角上看并不算庞大，但该商场对多源数据的整合和开发，不失为基于大数据管理的一种典型体现。

从理论上来说，每个企业都可能拥有大数据，但是并非每个企业都能够成为大数据企业。

大数据因其体量之“大”而得名，然而体量并非大数据的唯一特征，甚至也不是大数据最为重要的特征。巨大的体量凸显的是技术需求。而对于管理者而言，刻意追求巨大体量的数据并不具有多少现实意义，大数据更重要的特征在于其多样化的来源和形态、持续快速的产生和演变，以及对深度分析能力的高度依赖。因此，企业对大数据的驾驭和掌控，其核心并不在于拥有多大规模的数据，而在于是否能够对来自于企业内外部多样化信息源的涌流数据进行敏捷持续的捕捉和整合，并通过深度分析开发其商务价值。

在管理视角上，大数据既不是一种技术，也不是一种应用系统，而更应该是一种立足于企业内外部数据融合以提升管理效率、开拓价值创造模式的管理思维。

企业内部数据有两个主要维度：

一是与业务功能及流程紧密相关的数据，如库存信息、物料需求信息、生产计划信息、采购信息等，可统称为业务流程信息；

二是企业内员工及各种管理系统在其日常工作及活动中所创造、记录、交换和积累的信息，例如员工间的交流记录、工作心得、经验分享、活动新闻等，可统称为知识及沟通信息。

这两个数据维度的发展和融合，催生出了企业内部大数据，如图1.8所示。

在集成化企业系统、内部社交媒体以及深度数据分析技术的共同支撑下，杰克·韦尔奇所畅想的“无边界组织”在新兴环境下成为可能，并被赋予了新的内涵。部门边界、层级边界被紧密的业务联系和广泛的社交联系所弱化，结构化的业务流程信息与非结构化的知识及管理活动信息被多维度融合的深度数据分析能力连接在一起，从而使企业真正具有驾驭内部大数据的能力。

1.8.1　驾驭企业外部大数据

在企业外部的视角上，数据资源也包括两个维度：

一是与上下游交易直接相关的供应链信息，如交易报价信息、订单信息、上下游企业



图 1.8 企业内部大数据

库存及生产能力信息等；

二是市场及社会环境信息，如原材料价格走势、市场需求及消费者偏好信息、顾客服务及满意度信息等。

企业外部大数据的基本特征，也正是在这两个维度的发展之中呈现出来的。

供应链信息集成与社会化商务信息的融合，构成企业外部大数据的核心特征。来自于社交媒体信息源的市场环境信息与来自于组织间信息系统的供应链信息相结合，借助于深度数据分析技术实现面向企业商务网络的预测与优化，并支撑起实时化、精确化、个性化的消费者洞察与敏捷响应，在此基础上为基于网络协同及社会化商务的模式创新提供了丰富的可能性。因此，对外部大数据的管理和驾驭，也将成为现代企业在网络化的商务生态系统中占据主导地位并获取经营优势的关键途径。

1.8.2 成为“大数据企业”

基于以上分析，企业内部大数据的焦点，在于业务流程信息与知识及沟通信息的融合；企业外部大数据的焦点，在于供应链信息与市场及社会环境信息的融合。进而，大数据时代企业组织的基本内涵，在于内部大数据与外部大数据的全方位融合。如图 1.9 所示，大数据企业立足于内外部业务与社交媒体数据的集成交汇。

在这四大类型的数据之间，致力于大数据管理的企业可以有两种不同的发展策略。

第一种是以社交媒体与业务数据的融合为主导，以期通过敏捷响应快速发现并应对内外部环境中的变化和机遇。在这种策略下，面向高速数据流的实时数据采集和分析方法，将成为大数据管理的主要支撑手段。

第二种策略是以内外部数据融合为主导，以期通过全面汇集内外部信息，对中长期发展趋势做出准确的预判，从而实现高度优化的业务决策，并通过对信息环境的掌控，获取企业网络生态系统中的领导地位。在这种策略下，大规模多源异构数据的采集、清洗和整

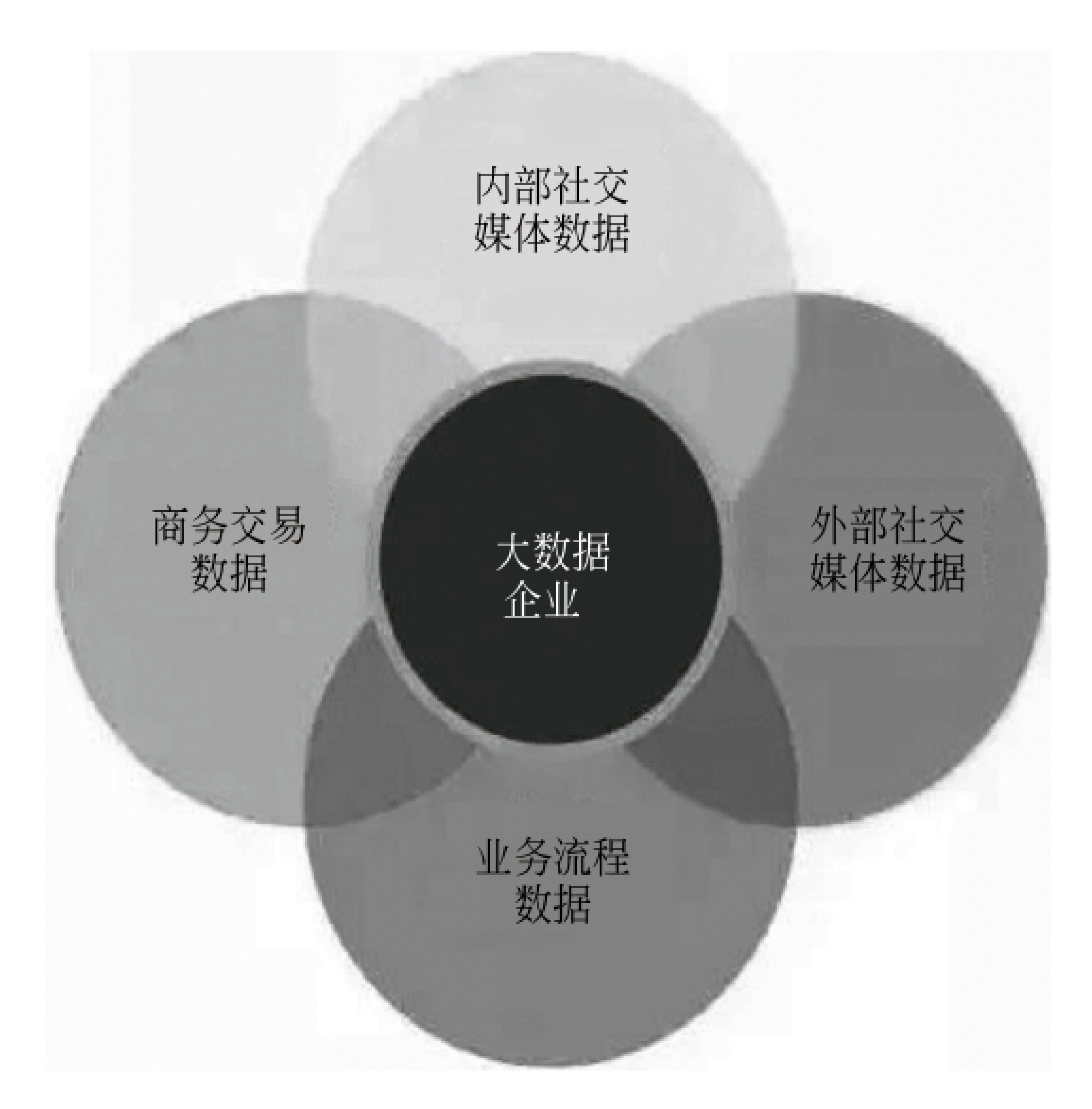


图 1.9　大数据企业的内外融合

合方法,将成为大数据管理的核心支撑。

1.8.3　如何挖掘企业大数据的价值

企业大数据的价值开发高度依赖于深度数据分析能力。从内外部融合的视角上看来,企业大数据分析包括三个基本维度,即内容、关系和时空。

1. 内容维度指的是数据本身所承载的信息内容

例如,G 公司是一家大型电信服务商,其内部建设实施了一套“班组博客”系统。在这个内部社交媒体平台上,公司中的 3000 多个工作团队都开设了自己的博客,用于发布和交流工作经验、生活体验等方面的内容。经过数年的发展,整个博客系统中积累了博文 700 多万篇,评论超过 1500 万条,并保持着每月 15 万篇以上的博文发表数量,年阅读量超过 1000 万篇次。

对于这一平台所积累的大量数据的价值开发,首先体现在对其信息内容的提炼上。平台上与工作相关的博文内容,如客服案例、经验分享等,经自动筛选分类、主题识别、关键词索引之后,被构建成企业知识库,为业务及管理工作提供快速有效的知识支撑,同时成为员工培训和自学的有力工具。而大量与工作无关的博文和评论内容,包括生活常识、娱乐信息、心情表达、心灵鸡汤等,在智能化的分类整理之后,也成为该公司的一个独特的文化情景,支撑着企业中活跃的氛围,强化了员工的文化认同。

2. 关系维度指的是数据及其所指代的对象之间的联系

在 G 公司的班组博客中,员工的发表、阅读、评论、回复、关注等行为详尽地反映了其相互之间密集而持续的联系,而这些联系毫无遗漏地被记录在平台的数据库之中。通过对这些关系结构的深度分析和挖掘,G 公司获得了对员工及团队的影响力、凝聚力、创造力的更为准确而深入的评估手段。进一步而言,博客平台的行为记录数据与业务系统中

的事务处理记录数据，以及员工及团队的绩效表现数据，也能够被有效地关联起来，从而使得管理者拥有强有力的工具，帮助其发现和理解员工的行为特质、工作表现、业务能力之间的潜在关联，进而实现良性优化的人员配置和人才培养。

时空维度指的是数据生成及传播的位置以及数据随时间演变的模式。对 G 公司而言，其数以千计的业务场所分散在众多城市的不同地点，因此，数据中的位置信息对于虚拟化的团队协同而言具有直接的意义。此外，位置信息也包括了数据在组织功能结构和层级结构中所处的位置。同时，在 G 公司的班组博客中，对特定话题时间演变规律的分析，也为管理者提供了有效的参考。其中对企业重要活动、运营理念相关信息在班组博客中的传播演变模式的跟踪，有效地揭示了员工对管理理念的认知、态度和接受过程。

3. 更深入的价值开发来自于上述三个维度的交叉综合

例如，内容维度与关系维度的结合，使得 G 公司能够识别员工的兴趣偏好、社交特质、工作性质以及工作表现之间的匹配关系，也能够更为准确地发现那些分散在不同的员工手中、但具有重要潜在影响力的经验、创意以及机遇信号。内容维度、关系维度与时空维度的结合，使得企业能够更为深入地理解不同的员工特质、知识技能、团队特性、热点偏好在整个组织中的分布，以及这些结构随时间演变的过程和趋势，从而更为有效地调度和配置这些资源。

这些维度上的分析需求，主要需要三方面的数据分析技术予以支撑。

第一类是全局视图技术。对于管理者而言，对大数据内容全局状况的把握，往往是开发大数据价值的一个基本需求。然而大数据的体量和结构复杂性往往远远超出人类认知的信息承载能力。因此，有效的技术应当能够在大量数据中提取出一个足够小的集合以呈现给管理者，并使得这个小集合能够充分地代表数据全局。例如，在 G 公司的博客平台上，一种“代表性博文提取”技术能够在每天所出现的数以千计的博文中自动选择出 10 篇。这 10 篇博文在很大程度上全面代表了当天所出现的数千篇文章，既充分反映热点，也不会忽略冷门信号，从而使得管理者能够通过阅读这些文章来了解全局。

第二类支撑技术是关联发现技术，其目标在于敏锐识别数据间的联系。例如，当 G 公司试图整合博客平台、业务系统、人力资源系统中的数据以全方位分析员工、团队特质以及绩效信息时，大量的数据属性之间所构成的复杂潜在关联网络，就需要强有力的关联发现技术来加以处理。

第三类支撑技术是动态跟踪技术，即实时化的流数据分析处理、快速增量数据分析。三方面技术都处于快速发展之中，但尚未全面成熟，有待于学界和业界的持续努力和探索。

1.8.4 大数据实质上是一种管理思维

从一定意义上说来，业务资源集成与社交媒体相融合的过程，是一个“信息去中心化”的过程。信息资源的创造和管理，从以往以经营和运作为核心的中心化模式，转化为以分散创造、自由传播、灵活汇聚为特征的众创模式。另一方面，内外部数据融合的过程，是一个“信息去边界化”的过程。企业部门之间的信息交换、企业之间的信息交换以及企业与市场环境的信息，以日益多样化、实时化的方式实现。

这样的转变对于企业组织及其员工而言，其影响将会是多方面的。正面的影响可能包括创新意识与创新行为的出现、员工能力和技能的发展、沟通满意度的提升、员工关系资本的建立和积累、员工对组织的认同和归属感的增加；而负面的影响则可能包括员工注意力分散、过度争论，以及负面情绪的传播等。所以，建设“大数据企业”的过程，也将会是一个伴随着困难与风险的过程。在此过程中，需要管理者有效地把握创新发展的长期收益与短期业绩之间的平衡，在推进大数据融合的同时防范和控制其中的组织风险，并审慎地思考和重新定义组织内外部边界。

换言之，对企业而言，大数据实质上是一种管理思维，其支点在于业务信息资源与社交媒体的融合，以及内外部数据的融合，在这样的支点上反思企业的组织形态、运作范式和价值创造模式，是“大数据企业”的真正内涵所在。

1.9　大数据应用案例之：男女嘉宾《非诚勿扰》牵手数据分析

《非诚勿扰》是由中国大陆江苏卫视制作的一档以婚恋交友为核心的社会生活服务真人秀节目，于 2010 年 1 月 15 日开播，节目内容取材自在全世界范围被广泛采用的英国独立电视台的两性联谊节目 *Take Me Out*，和 2008—2009 年播出的澳大利亚节目 *Taken Out*。自开播以来，《非诚勿扰》收视率在中国大陆各个卫星电视节目中名列前茅，且收视率日渐攀升。由江苏电视台新闻节目主持人孟非主持，现另由黄菡分析点评，孟非、黄磊、刘烨、宁财神、曾子航等均担任过男嘉宾。《非诚勿扰》节目页面见图 1.10。



图 1.10　江苏卫视制作的《非诚勿扰》节目

该节目如此火爆的收视率和普及度，使其男女嘉宾备受关注。比如女嘉宾身份问题、男嘉宾“托儿”、炒作等问题，也成为大家八卦的主题。截止到 2015 年第三季度，一共做了 539 期节目，至少 1508 名女嘉宾和 2382 名男嘉宾参与节目，成功促成了其中 419 对牵手男女嘉宾！其牵手成功页面如图 1.11 所示。



图 1.11　剑桥大数据博士上非诚勿扰，管女嘉宾叫样本，牵手成功!!!

较熟悉的观众都知道女神位这个词儿。通过节目录制现场的示意图，最中间 11～14 号女嘉宾是正对男嘉宾的，似乎话题和曝光率都颇高也备受关注。那么她们既然是女神了，是不是能尽快获得自己的男神呢？通过对牵手男女嘉宾的分析我们发现，真正的牵手女神却在 20 号位置左右！一共产生过 57 对牵手女嘉宾，几乎是 11～14 号位置牵手女嘉宾的总和了！我们也在分析，是不是站到了女神位，由于心理上的变化而使得男嘉宾被更多灭灯呢？

女神位置大数据分析如图 1.12 所示。

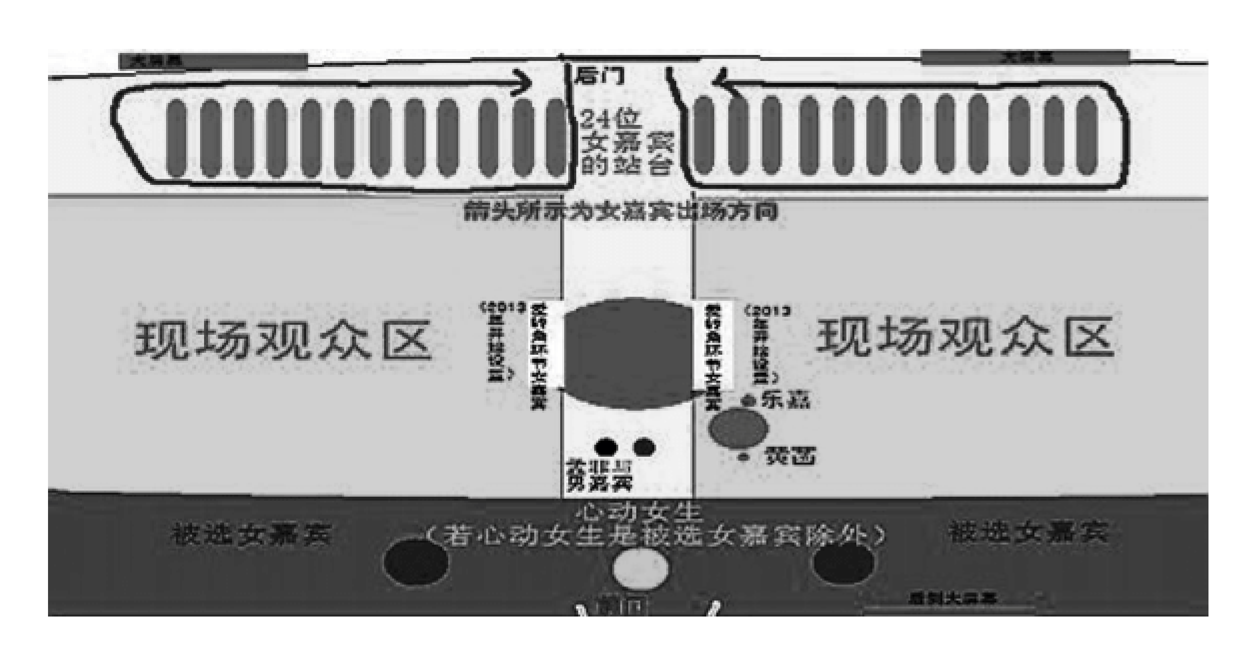


图 1.12　女神位置大数据分析

而对男嘉宾我们发现，最容易牵手的出场位置是 4 号出场！而 1 号往往最容易当炮灰。看来老话说得好——万事开头难。大多数节目都是 5 个男嘉宾，如果第四个男嘉宾没有牵手成功，恐怕第五个押宝牵手的几率也不会特别高，那么各位白富美就要再多站一期节目才能获得符合自己心意的高富帅了！是不是这种心理上的变化也使得 4 号男嘉宾更容易成功呢！如果你参加节目能够在此位置出场可要好好把握了哦，大数据告诉你牵手概率不低哦！

那么什么样的男女嘉宾比较受欢迎呢？我们通过男女嘉宾的地理位置、职业、年龄和牵手比率几个维度，发现中国的女嘉宾和欧美的男嘉宾比较受欢迎，其中自由职业、教师、企业职员比较受欢迎，私营业主身份的男嘉宾则最受欢迎。

牵手分析：中国的女嘉宾比较受欢迎如图 1.13 所示。

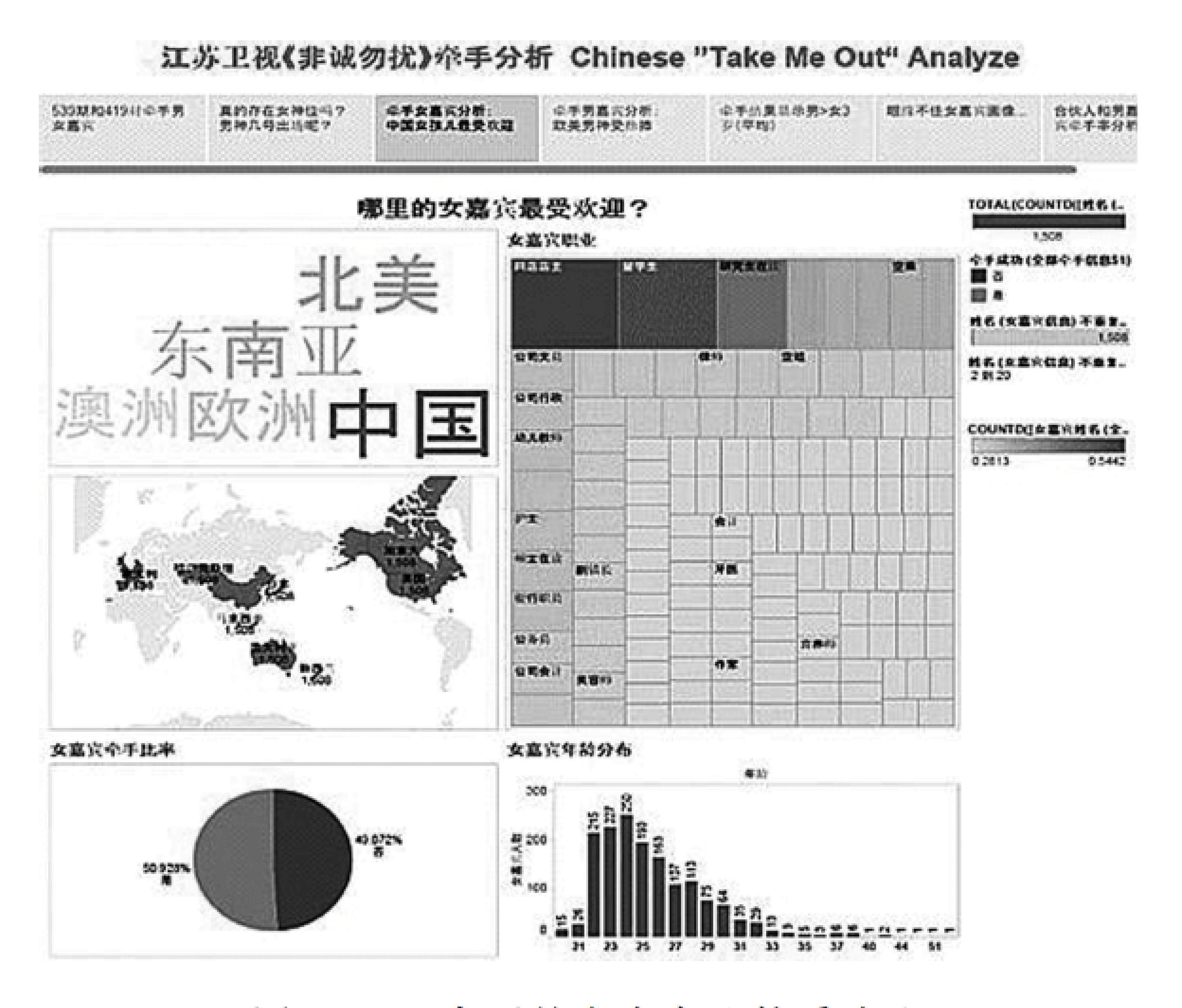


图 1.13　中国的女嘉宾比较受欢迎

牵手分析：欧美的男嘉宾比较受欢迎如图 1.14 所示。

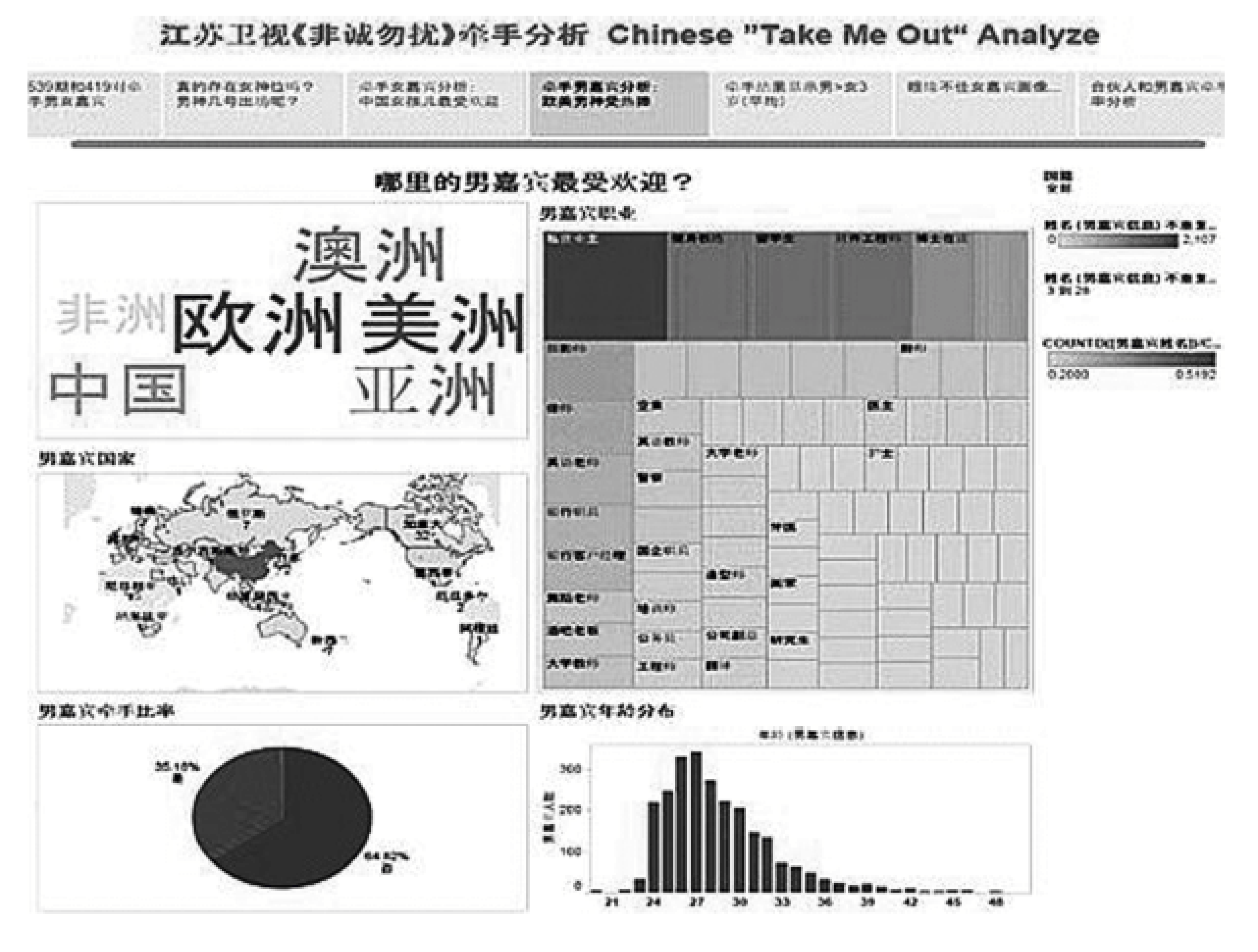


图 1.14　欧美的男嘉宾比较受欢迎

如果从年龄分布来看，男嘉宾普遍比女嘉宾的年龄要大。牵手男嘉宾的年龄段集中在 24～31 岁，女嘉宾则集中在 22～25 岁。而这种年龄分布于当下社会新组建家庭的物质需求也比较符合。可见来的男女嘉宾都是比较务实的，也符合了节目的主题——非诚勿扰。

牵手分析：牵手年龄分析如图 1.15 所示。

图 1.15　牵手年龄分析

在分析数据过程中我们还发现，有些女嘉宾可能确实眼缘不佳，在节目停留了 50 期以上也没能牵手成功。而眼缘不佳的女嘉宾集中分布在 25 岁以上。随着节目的改制这种现象在 2014 年以后逐步减少了，男嘉宾看到这个消息应该拍手称快了！

既然作为一款真人秀节目，就一定少不了娱乐的成分。根据数据显示，刘烨、曾子航、刘恺威三位男嘉宾在台上的时候，男嘉宾的牵手几率比较高，达到了 44%以上，他们真正做到了男嘉宾的好帮衬！而于正老师在台上的时候牵手率只有 26%。

其实还有很多主题因为缺少有力的数据支持，所以没有得以实现。比如，心动女生画像分析、星座牵手几率、旅游奖励的有效性分析、男嘉宾灭灯分析。现在很多人相亲的时候，其实自己都不清楚自己到底想要找一个什么样的牵手对象。根据上述分析，可以比较客观地知道什么样的男嘉宾更适合台上的女嘉宾。可以想象，如果男嘉宾上台的时候，女嘉宾手里也有一个数字，表示根据大数据计算这个男嘉宾和你的契合度是多少，那会是什么结果。或许大数据的判断比女嘉宾更懂你自己呢！

习题与思考题

一、选择题

1. 大数据的 4V 特点：Volume、Velocity、Variety、Veracity，它们的含义分别是（　　）、（　　）、（　　）、（　　）。

A. 价值密度低　　B. 处理速度快

C. 数据类型繁多　　D. 数据体量巨大

2. 大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行（　　）。

A. 数据信息　　B. 专业化处理

C. 速度处理　　D. 内容处理

3. 尿布啤酒案例是大数据分析的（　　）。

A. A/B 测试　　B. 分类

C. 关联规则挖掘　　D. 数据聚类

4. 当前大数据技术的基础是由（　　）首先提出的。

A. 微软　　B. 百度

C. Google　　D. 阿里巴巴

5. 根据不同的业务需求来建立数据模型，抽取最有意义的向量，决定选取哪种方法的数据分析角色人员是（　　）。

A. 数据管理人员　　B. 数据分析员

C. 研究科学家　　D. 软件开发工程师

6. 智慧城市的构建，不包含（　　）。

A. 数字城市　　B. 物联网

C. 联网监控　　D. 云计算

7. 大数据的最显著特征是（　　）。

A. 数据规模大　　B. 数据类型多样
C. 数据处理速度快　　D. 数据价值密度高

8. 大数据时代，数据使用的关键是(　　)。
A. 数据收集　　B. 数据存储
C. 数据分析　　D. 数据再利用

9. 支撑大数据业务的基础是(　　)。
A. 数据科学　　B. 数据应用
C. 数据硬件　　D. 数据人才

10. 大数据不是要教机器像人一样思考。相反，它是(　　)。
A. 把数学算法运用到海量的数据上来预测事情发生的可能性
B. 被视为人工智能的一部分
C. 被视为一种机器学习
D. 预测与惩罚

11. 大数据的发展，使信息技术变革的重点从关注技术转向关注(　　)。
A. 信息　　B. 数字　　C. 文字　　D. 方位

二、问答题

1. 简述大数据的定义和特点。
2. 大数据的社会价值体现在哪些方面?
3. 简述商业大数据的类型和价值挖掘方法。
4. 基于大数据分析的商业模式创新有哪些?
5. 如何成为“大数据企业”?



大数据技术

第2章　基础架构——云计算平台

2.1　大数据处理的基础架构

人们研究大数据，或是利用大数据技术，其战略意义并不在于是谁掌握了多么庞大的大数据信息，而是在于谁能否将已经捕捉到的那些含有一定意义的数据通过专业化处理，将其变成一种数据信息资产。这也是大数据分析的真正目的。

谁都不能否认，大数据既是一种科技，也是一种资产。既然大数据是一种资产，那么，如何利用大数据这种资产最终实现盈利，才是运用大数据的关键。可是，将大数据加工成有增值的数据，并不是一件轻而易举的事情。

第一，研究大数据绝对离不开计算机的云计算技术。

从某种观点上看，没有计算机的云计算技术，就不会有大数据的被分析和利用。大数据技术跟计算机云计算技术的关系就像是一只手的手心和手背，是密不可分的。因为分析和处理大数据是无法用某一台计算机来完成的，它必须采用计算机的分布式架构，处理大数据的特色就是在于对那些海量性的数据进行分布式的数据挖掘，但这种分布式的大数据挖掘，还必须依托计算机的分布式处理，因为计算机的分布式数据库或是云存储以及计算机中的虚拟化技术，可以支撑起对大数据相关技术处理的能力。

第二，计算机云计算技术时代的到来将大数据处理变为了现实。

大数据内部所含有的资产性质，被计算机云计算技术得到了实实在在的验证，由此而引出来的效果，就是让很多人都对大数据有了更多的关注或是重视。可用大数据来形容某家公司所创造的那些大量非结构化数据和半结构化数据，但不能将这些数据下载到关系型的数据库中进行处理，因为这样会在分析数据中浪费较多的时间或金钱。大数据的分析必须要跟计算机的云计算技术紧密连在一起，只有这样，才能将大数据的价值变成资产性的价值，并将大数据处理真正变成一种现实。

2.2　云计算网络

云计算(Cloud Computing)是分布式计算技术的一种，其最基本的概念，是通过网络将庞大的计算处理程序自动分拆成无数个较小的子程序，再交由多部服务器所组成的庞大系统经搜寻、计算分析之后将处理结果回传给用户。以前的大规模分布式计算技术即为“云计算”的概念起源。

云计算的核心思想，是将大量用网络连接的计算资源统一管理和调度，构成一个计算资源池向用户按需服务。云计算的一个核心理念就是通过不断提高“云”的处理能力，进

而减少用户终端的处理负担,最终使用户终端简化成一个单纯的输入输出设备,并能按需享受“云”的强大计算处理能力!

2.2.1 云计算简介

1. 简介

云计算是网格计算(Grid Computing)、分布式计算(Distributed Computing)、并行计算(Parallel Computing)、效用计算(Utility Computing)、网络存储(Network Storage Technologies)、虚拟化(Virtualization)、负载均衡(Load Balance)等传统计算机技术和网络技术发展融合的产物。

它旨在通过网络把多个成本相对较低的计算实体整合成一个具有强大计算能力的完美系统,并借助 SaaS、PaaS、IaaS、MSP 等先进的商业模式把这强大的计算能力分布到终端用户手中。云计算将所有的计算资源集中起来,并由软件实现自动管理,无须人为参与。这使得应用提供者无须为烦琐的细节而烦恼,能够更加专注于自己的业务,有利于创新和降低成本。

2. 定义

1) 狭义云计算

提供资源的网络被称为“云”。“云”中的资源在使用者看来是可以无限扩展的,并且可以随时获取,按需使用,随时扩展,按使用付费。这种特性经常被称为像水电一样使用 IT 基础设施。

2) 广义云计算

这种服务可以是 IT 和软件、互联网相关的,也可以是任意其他的服务。这种资源池称为“云”。“云”是一些可以自我维护和管理的虚拟计算资源,通常为一些大型服务器集群,包括计算服务器、存储服务器、宽带资源等等。

云计算是并行计算、分布式计算和网格计算的发展,或者说是这些计算机科学概念的商业实现。云计算是虚拟化、效用计算、IaaS(基础设施即服务)、PaaS(平台即服务)、SaaS(软件即服务)等概念混合演进并跃升的结果。总的来说,云计算可以算作是网格计算的一个商业演化版。

3. 原理

云计算的基本原理是,通过使计算分布在大量的分布式计算机上,而非本地计算机或远程服务器中,企业数据中心的运行将与互联网更相似。这使得企业能够将资源切换到需要的应用上,根据需求访问计算机和存储系统。

这可是一种革命性的举措,打个比方,这就好比是从古老的单台发电机模式转向了电厂集中供电的模式。它意味着计算能力也可以作为一种商品进行流通,就像煤气、水电一样,取用方便,费用低廉。其最大的不同在于,它是通过互联网进行传输的。

在未来,只需要一台笔记本或者一个手机,就可以通过网络服务来实现我们需要的一切,甚至包括超级计算这样的任务。从这个角度而言,最终用户才是云计算的真正拥有者。云计算的应用包含这样的一种思想,把力量联合起来,给其中的每一个成员使用。

4. 特点

1）数据安全可靠

首先，云计算提供了最可靠、最安全的数据存储中心，用户不用再担心数据丢失、病毒入侵等麻烦。

多人觉得数据只有保存在自己看得见、摸得着的电脑里才最安全，其实不然。你的电脑可能会因为自己不小心而被损坏，或者被病毒攻击，导致硬盘上的数据无法恢复，而有机会接触你的电脑的不法之徒则可能利用各种机会窃取你的数据。此前轰动一时的“艳照门”事件据报道不也是因为电脑送修而造成个人数据外泄的吗？

反之，当你的文档保存在类似 Google Docs 的网络服务上，当你把自己的照片上传到类似 Google Picasa Web 的网络相册里，你就再也不用担心数据的丢失或损坏。因为在“云”的另一端，有全世界最专业的团队来帮你管理信息，有全世界最先进的数据中心来帮你保存数据。同时，严格的权限管理策略可以帮助你放心地与你指定的人共享数据。这样，你不用花钱就可以享受到最好、最安全的服务，甚至比在银行里存钱还方便。

2）客户端需求低

其次，云计算对用户端的设备要求最低，使用起来也最方便。

大家都有过维护个人电脑上种类繁多的应用软件的经历。为了使用某个最新的操作系统，或使用某个软件的最新版本，我们必须不断升级自己的电脑硬件。为了打开朋友发来的某种格式的文档，我们不得不疯狂寻找并下载某个应用软件。为了防止在下载时引入病毒，我们不得不反复安装杀毒和防火墙软件。所有这些麻烦事加在一起，对于一个刚刚接触计算机、刚刚接触网络的新手来说不啻一场噩梦！

如果你再也无法忍受这样的电脑使用体验，云计算也许是你的最好选择。你只要有一台可以上网的电脑，有一个你喜欢的浏览器，你要做的就是在浏览器中输入 URL，然后尽情享受云计算带给你的无限乐趣。

你可以在浏览器中直接编辑存储在“云”的另一端的文档，你可以随时与朋友分享信息，再也不用担心你的软件是否是最新版本，再也不用为软件或文档染上病毒而发愁。因为在“云”的另一端，有专业的 IT 人员帮你维护硬件，帮你安装和升级软件，帮你防范病毒和各类网络攻击，帮你做你以前在个人电脑上所做的一切。

3）轻松共享数据

此外，云计算可以轻松实现不同设备间的数据与应用共享。

大家不妨回想一下，你自己的联系人信息是如何保存的。一个最常见的情形是，你的手机里存储了几百个联系人的电话号码，你的个人电脑或笔记本电脑里则存储了几百个电子邮件地址。为了方便在出差时发邮件，你不得不在个人电脑和笔记本电脑之间定期同步联系人信息。买了新的手机后，你不得不在旧手机和新手机之间同步电话号码。对了，还有你的 PDA 以及你办公室里的电脑。

考虑到不同设备的数据同步方法种类繁多，操作复杂，要在这许多不同的设备之间保存和维护最新的一份联系人信息，你必须为此付出难以计数的时间和精力。这时，你需要

用云计算来让一切都变得更简单。在云计算的网络应用模式中,数据只有一份,保存在“云”的另一端,你的所有电子设备只需要连接互联网,就可以同时访问和使用同一份数据。

仍然以联系人信息的管理为例,当你使用网络服务来管理所有联系人的信息后,你可以在任何地方用任何一台电脑找到某个朋友的电子邮件地址,可以在任何一部手机上直接拨通朋友的电话号码,也可以把某个联系人的电子名片快速分享给好几个朋友。当然,这一切都是在严格的安全管理机制下进行的,只有对数据拥有访问权限的人,才可以使用或与他人分享这份数据。

4）可能无限多

最后,云计算为我们使用网络提供了几乎无限多的可能,为存储和管理数据提供了几乎无限多的空间,也为我们完成各类应用提供了几乎无限强大的计算能力。想象一下,当你驾车出游的时候,只要用手机连入网络,就可以直接看到自己所在地区的卫星地图和实时的交通状况,可以快速查询自己预设的行车路线,可以请网络上的好友推荐附近最好的景区和餐馆,可以快速预订目的地的宾馆,还可以把自己刚刚拍摄的照片或视频剪辑分享给远方的亲友……

离开了云计算,单单使用个人电脑或手机上的客户端应用,我们是无法享受这些便捷服务的。个人电脑或其他电子设备不可能提供无限量的存储空间和计算能力,但在“云”的另一端,由数千台、数万台甚至更多服务器组成的庞大的集群却可以轻易地做到这一点。个人和单个设备的能力是有限的,但云计算的潜力却几乎是无限的。当你把最常用的数据和最重要的功能都放在“云”上时,我们相信,你对电脑、应用软件乃至网络的认识会有翻天覆地的变化,你的生活也会因此而改变。

互联网的精神实质是自由、平等和分享。作为一种最能体现互联网精神的计算模型,云计算必将在不远的将来展示出强大的生命力,并将从多个方面改变我们的工作和生活。无论是普通网络用户,还是企业员工;无论是 IT 管理者,还是软件开发人员,他们都能亲身体验到这种改变。

5）营销

通过网络,把多个成本较低的计算实体,整合成一个具有强大营销能力的完美系统。核心理念就是通过不断提高“云”的覆盖能力,以及“云”之间的逻辑计算能力,从而达到系统营销的结果,它可以减少用户的经济负担,最终使用户简化到只要在家里,通过一台终端,就可以得到近乎无限数量的优质客户,享受“营销云”带来的强大经济利益。

狭义云营销:帮客户销售产品,快速建立全国营销渠道,获取经济利益。

广义云营销:树立企业品牌形象,获取更多的社会资源等。

2.2.2 云计算系统的体系结构

1. 云计算逻辑结构

云计算平台是一个强大的“云”网络,连接了大量并发的网络计算和服务,可利用虚拟化技术扩展每一个服务器的能力,将各自的资源通过云计算平台结合起来,提供超级计算和存储能力。通用的云计算逻辑结构如图 2.1 所示。

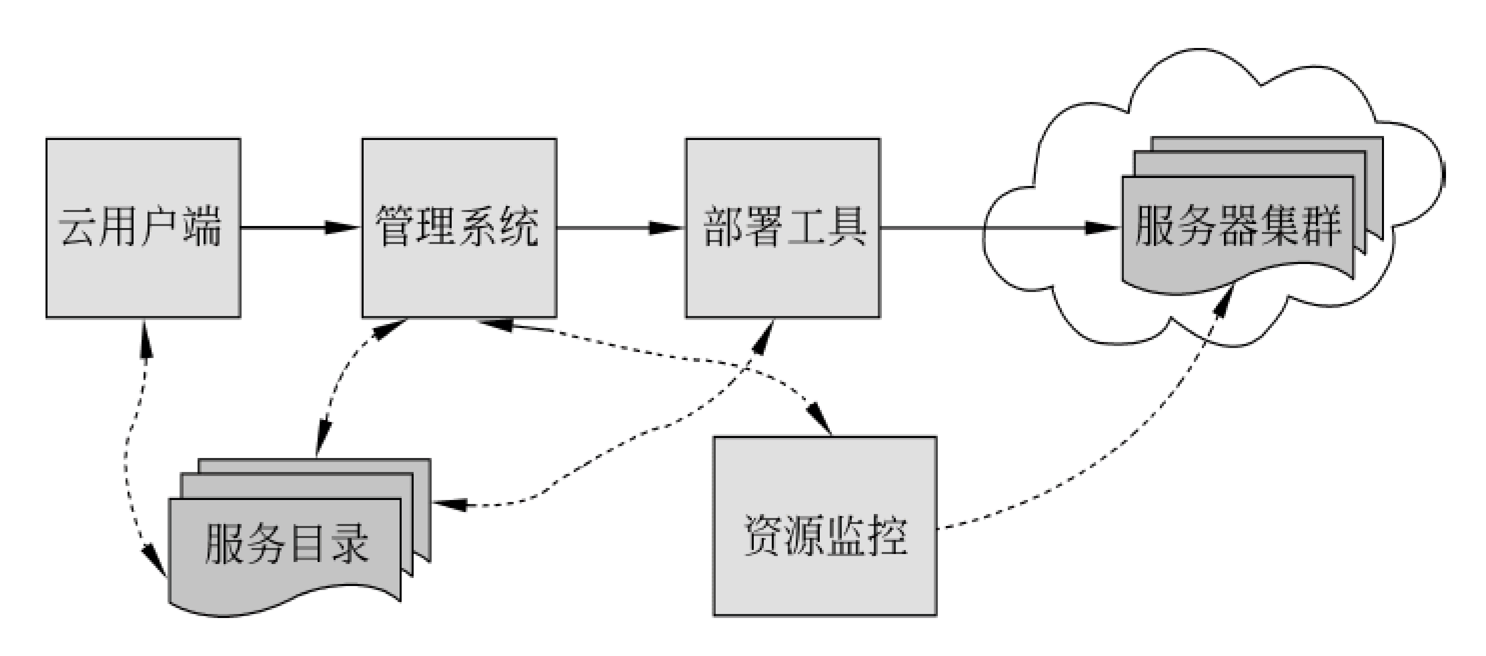


图 2.1　云计算逻辑结构

1）云用户端

提供云用户请求服务的交互界面，也是用户使用云的入口，用户通过 Web 浏览器可以注册、登录及定制服务、配置和管理用户。打开应用实例，就像在本地操作桌面系统一样。

2）服务目录

云用户在取得相应权限（付费或其他限制）后可以选择或定制的服务列表，也可以对已有服务进行退订操作，在云用户端界面生成相应的图标或列表的形式展示相关的服务。

3）管理系统和部署工具

提供管理和服务，能管理云用户，能对用户授权、认证、登录进行管理，并可以管理可用计算资源和服务，接收用户发送的请求，根据用户请求并转发到相应的相应程序，调度资源智能地部署资源和应用，动态地部署、配置和回收资源。

4）监控

监控和计量云系统资源的使用情况，以便做出迅速反应，完成结点同步配置、负载均衡配置和资源监控，确保资源能顺利分配给合适的用户。

5）服务器集群

虚拟的或物理的服务器，由管理系统管理，负责高并发量的用户请求处理、大运算量计算处理、用户 Web 应用服务，云数据存储时采用相应数据切割算法采用并行方式上传和下载大容量数据。

用户可通过云用户端从列表中选择所需的服务，其请求通过管理系统调度相应的资源，并通过部署工具分发请求、配置 Web 应用。

2. 云计算的主要服务形式

目前，云计算的主要服务形式有 IaaS(Software as a Service，基础设施即服务)、PaaS (Platform as a Service，平台即服务)、SaaS(Infrastructure as a Service，软件即服务)，如图 2.2 所示。

1）软件即服务(SaaS)

SaaS 服务提供商将应用软件统一部署在自己的服务器上，用户根据需求通过互联网向厂商订购应用软件服务，服务提供商根据客户所定软件的数量、时间的长短等因素收费，并且通过浏览器向客户提供软件的模式。

这种服务模式的优势是，由服务提供商维护和管理软件、提供软件运行的硬件设施，

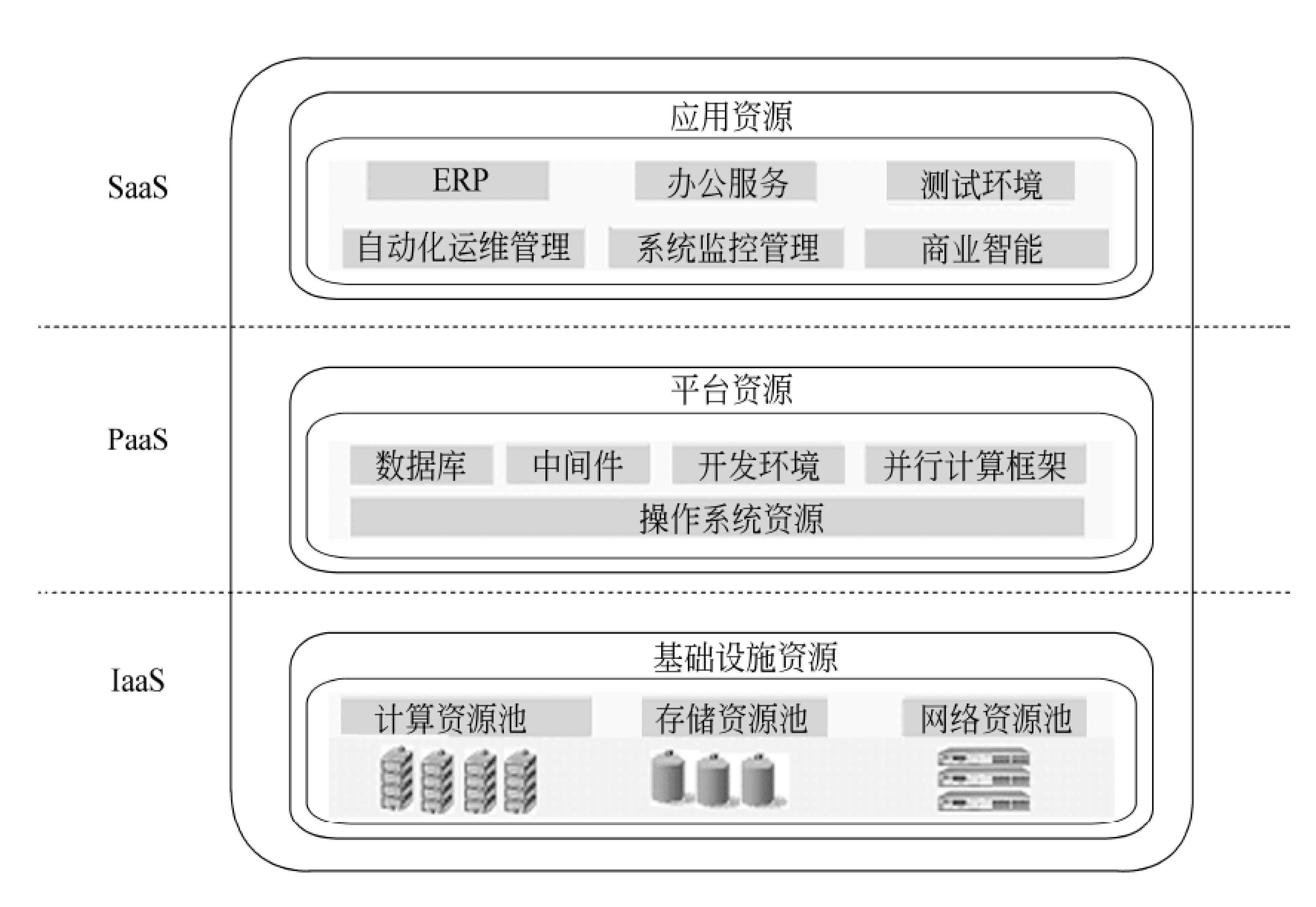


图 2.2 云计算的主要服务形式

用户只需拥有能够接入互联网的终端，即可随时随地使用软件。这种模式下，客户不再像传统模式那样资金在硬件、软件、维护人员方面花费大量，只需要支出一定的租赁服务费用，通过互联网就可以享受到相应的硬件、软件和维护服务，这是网络应用最具效益的营运模式。对于小型企业来说，SaaS 是采用先进技术的最好途径。

以企业管理软件来说，SaaS 模式的云计算 ERP 可以让客户根据并发用户数量、所用功能多少、数据存储容量、使用时间长短等因素不同组合按需支付服务费用，既不用支付软件许可费用，也不需要支付采购服务器等硬件设备费用，也不需要支付购买操作系统、数据库等平台软件费用，也不用承担软件项目定制、开发、实施费用，也不需要承担 IT 维护部门开支费用。实际上，云计算 ERP 正是继承了开源 ERP 免许可费用只收服务费用的最重要特征，是突出了服务的 ERP 产品。

目前，Salesforce.com 是提供这类服务最有名的公司，Google Doc、Google Apps 和 Zoho Office 也属于这类服务。

2）平台即服务(PaaS)

把开发环境作为一种服务来提供。这是一种分布式平台服务，厂商提供开发环境、服务器平台、硬件资源等服务给客户，用户在其平台基础上定制开发自己的应用程序并通过其服务器和互联网传递给其他客户。PaaS 能够给企业或个人提供研发的中间件平台，提供应用程序开发、数据库、应用服务器、试验、托管及应用服务。

Google App Engine，Salesforce 的 force.com 平台，八百客的 800APP 是 PaaS 的代表产品。以 Google App Engine 为例，它是一个由 Python 应用服务器群、BigTable 数据库及 GFS 组成的平台，为开发者提供一体化主机服务器及可自动升级的在线应用服务。用户编写应用程序并在 Google 的基础架构上运行就可以为互联网用户提供服务，Google 提供应用运行及维护所需要的平台资源。

3）基础设施服务（IaaS）

IaaS 即把厂商的由多台服务器组成的"云端"基础设施作为计量服务提供给客户。它将内存、I/O 设备、存储和计算能力整合成一个虚拟的资源池，为整个业界提供所需要的存储资源和虚拟化服务器等服务。这是一种托管型硬件方式，用户付费使用厂商的硬件设施。例如 Amazon Web 服务（AWS）、IBM 的 BlueCloud 等均是将基础设施作为服务出租。

IaaS 的优点是用户只需低成本硬件，按需租用相应计算能力和存储能力，大大降低了用户在硬件上的开销。

3. 云计算应用

目前，以 Google 云的应用最具代表性，例如 GoogleDocs、GoogleApps、Googlesites 以及云计算应用平台 GoogleApp Engine。

1）GoogleDocs

GoogleDocs 是最早推出的云计算应用，是软件即服务思想的典型应用。它是类似于微软的 Office 的在线办公软件。它可以处理和搜索文档、表格、幻灯片，并可以通过网络和他人分享并设置共享权限。Google 文件是基于网络的文字处理和电子表格程序，可提高协作效率，多名用户可同时在线更改文件，并可以实时看到其他成员所做的编辑操作。

用户只需一台接入互联网的计算机和可以使用 Google 文件的标准浏览器即可在线创建和管理、实时协作、权限管理、共享、搜索能力、修订历史记录功能，以及随时随地访问的特性，大大提高了文件操作的共享和协同能力。

2）GoogleAPPs

GoogleAPPs 是 Google 企业应用套件，使用户能够处理日渐庞大的信息量，随时随地保持联系，并可与其他同事、客户和合作伙伴进行沟通、共享和协作。它集成了 Cmail、GoogleTalk、Google 日历、GoogleDocs 以及最新推出的云应用 GoogleSites、API 扩展以及一些管理功能，包含了通信、协作与发布、管理服务三方面的应用，并且拥有着云计算的特性，能够更好地实现随时随地协同共享。另外，它还具有低成本的优势和托管的便捷性，用户无须自己维护和管理搭建的协同共享平台。

3）Googlesites

Googlesites 是 Google 最新发布的云计算应用，作为 GoogleAPPs 的一个组件出现。它是一个侧重于团队协作的网站编辑工具，可利用它创建一个各种类型的团队网站，通过 Googlesites 可将所有类型的文件包括文档、视频、相片、日历及附件等与好友、团队或整个网络分享。

4）Google AppEngine

Google AppEngine 是 Google 在 2008 年 4 月发布的一个平台，使用户可以在 Google 的基础架构上开发和部署运行自己的应用程序。目前，Google AppEngine 支持 Python 语言和 Java 语言，每个 Google AppEngine 应用程序可以使用达到 500MB 的持久存储空间及可支持每月 500 万综合浏览量的带宽和 CPU。并且，Google AppEngine 应用程序易于构建和维护，并可根据用户的访问量和数据存储需要的增长轻松扩展。

同时,用户的应用可以和 Google 的应用程序集成,Google AppEngine 还推出了软件开发套件(SDK),包括可以在用户本地计算机上模拟所有 Google AppEngine 服务的网络服务器应用程序。

4. 云计算技术体系结构

由于云计算分为 IaaS、PaaS 和 SaaS 三种类型,不同的厂家又提供了不同的解决方案,目前还没有一个统一的技术体系结构;综合不同厂家的方案,给出一个供商榷的云计算技术体系结构。这个体系结构如图 2.3 所示,它概括了不同解决方案的主要特征。

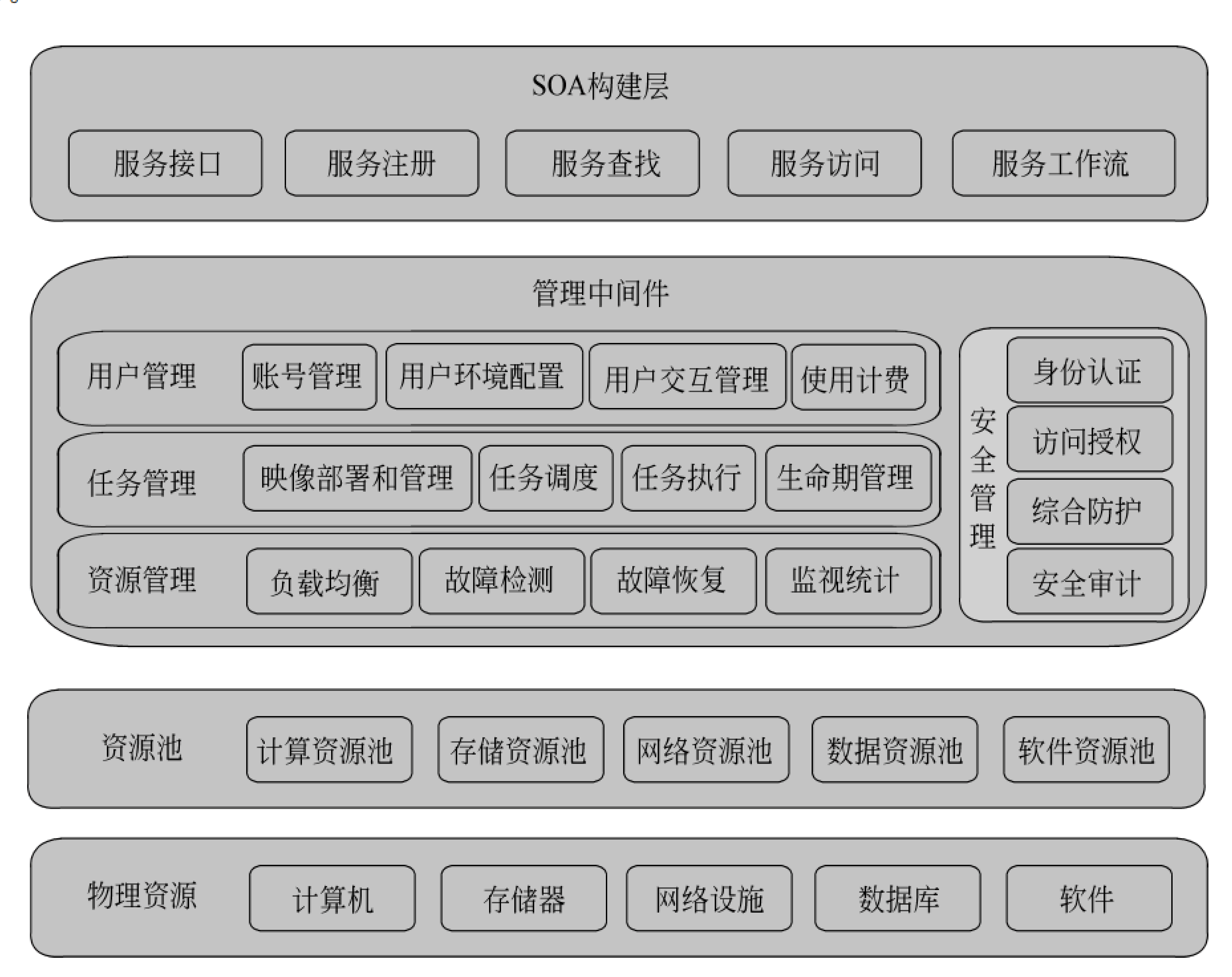


图 2.3 云计算体系结构

云计算技术体系结构分为 4 层:物理资源层、资源池层、管理中间件层和 SOA 构建层。

(1) 物理资源层包括计算机、存储器、网络设施、数据库和软件等;

(2) 资源池层是将大量相同类型的资源构成同构或接近同构的资源池,如计算资源池、数据资源池等。构建资源池更多是物理资源的集成和管理工作,例如研究在一个标准集装箱的空间如何装下 2000 个服务器、解决散热和故障结点替换的问题并降低能耗。

(3) 管理中间件层负责对云计算的资源进行管理,并对众多应用任务进行调度,使资源能够高效、安全地为应用提供服务;

(4) SOA 构建层将云计算能力封装成标准的 Web Services 服务,并纳入到 SOA 体系进行管理和使用,包括服务注册、查找、访问和构建服务工作流等。管理中间件和资源池层是云计算技术的最关键部分,SOA 构建层的功能更多依靠外部设施提供。

5. 云计算简化实现机制

基于上述体系结构，以 IaaS 云计算为例，简述云计算的实现机制，如图 2.4 所示。

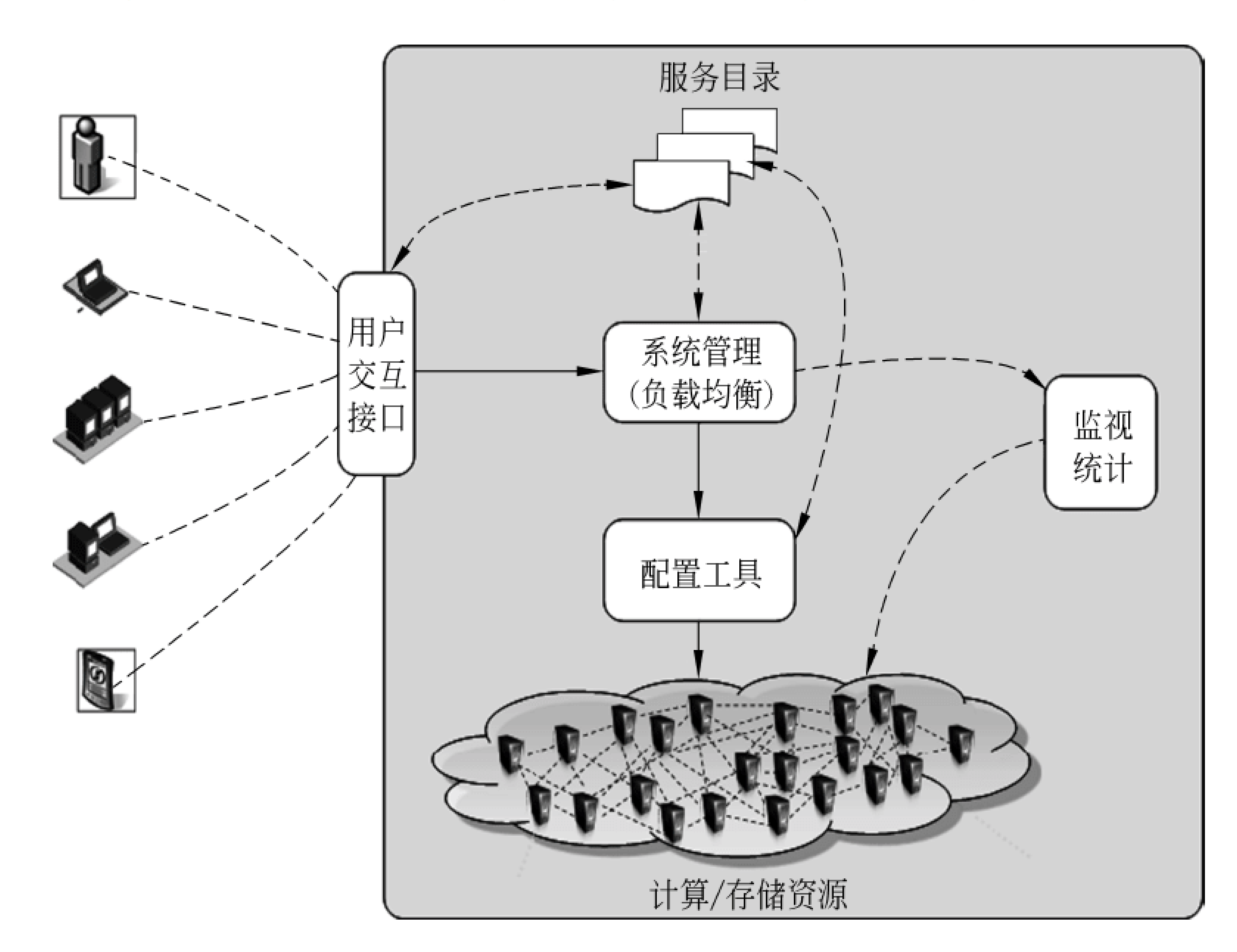


图 2.4　云计算简化实现机制

（1）用户交互接口：向应用以 Web Services 方式提供访问接口，获取用户需求。

（2）服务目录：是用户可以访问的服务清单。系统管理模块负责管理和分配所有可用的资源，其核心是负载均衡。配置工具负责在分配的结点上准备任务运行环境。

（3）监视统计模块：负责监视结点的运行状态，并完成用户使用结点情况的统计。执行过程并不复杂。

（4）用户交互接口：允许用户从目录中选取并调用一个服务。该请求传递给系统管理模块后，它将为用户分配恰当的资源，然后调用配置工具来为用户准备运行环境。

2.2.3　云计算服务层次

1. 云计算服务层次

在云计算中，根据其服务集合所提供的服务类型，整个云计算服务集合被划分成 4 个层次：应用层、平台层、基础设施层和虚拟化层。这 4 个层次每一层都对应着一个子服务集合，为云计算服务层次模型如图 2.5 所示。

云计算的服务层次是根据服务类型即服务集合来划分，与大家熟悉的计算机网络体系结构中层次的划分不同。在计算机网络中每个层次都实现一定的功能，层与层之间有一定关联。而云计算体系结构中的层次是可以分割的，即某一层次可以单独完成一项用户的请求而不需要其他层次为其提供必要的服务和支持。

在云计算服务体系结构中各层次与相关云产品对应。

（1）应用层对应 SaaS 软件即服务，如 GoogleApps、SoftWare＋Services；

（2）平台层对应 PaaS 平台即服务，如 IBM IT Factory、Google APPEngine、

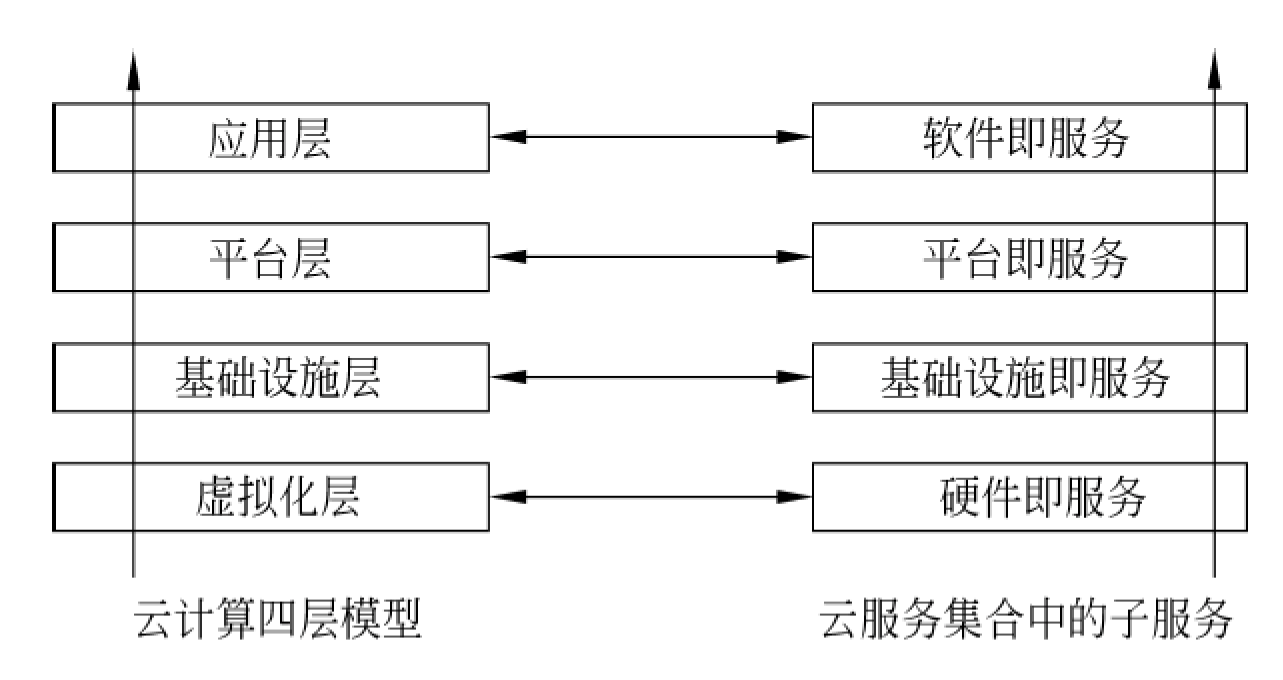


图 2.5　云计算服务层次模型

Force.com；

(3) 基础设施层对应 IaaS 基础设施即服务，如 Amazo Ec2、IBM Blue Cloud、Sun Grid；

(4) 虚拟化层对应硬件即服务，结合 Paas 提供硬件服务，包括服务器集群及硬件检测等服务。

大部分的云计算基础构架是由通过数据中心传送的可信赖的服务和创建在服务器上的不同层次的虚拟化技术组成的。人们可以在任何有提供网络基础设施的地方使用这些服务。“云”通常表现为对所有用户的计算需求的单一访问点。人们通常希望商业化的产品能够满足服务质量(QoS)的要求，并且一般情况下要提供服务水平协议。开放标准对于云计算的发展是至关重要的，并且开源软件已经为众多的云计算实例提供了基础，如图 2.6 所示。

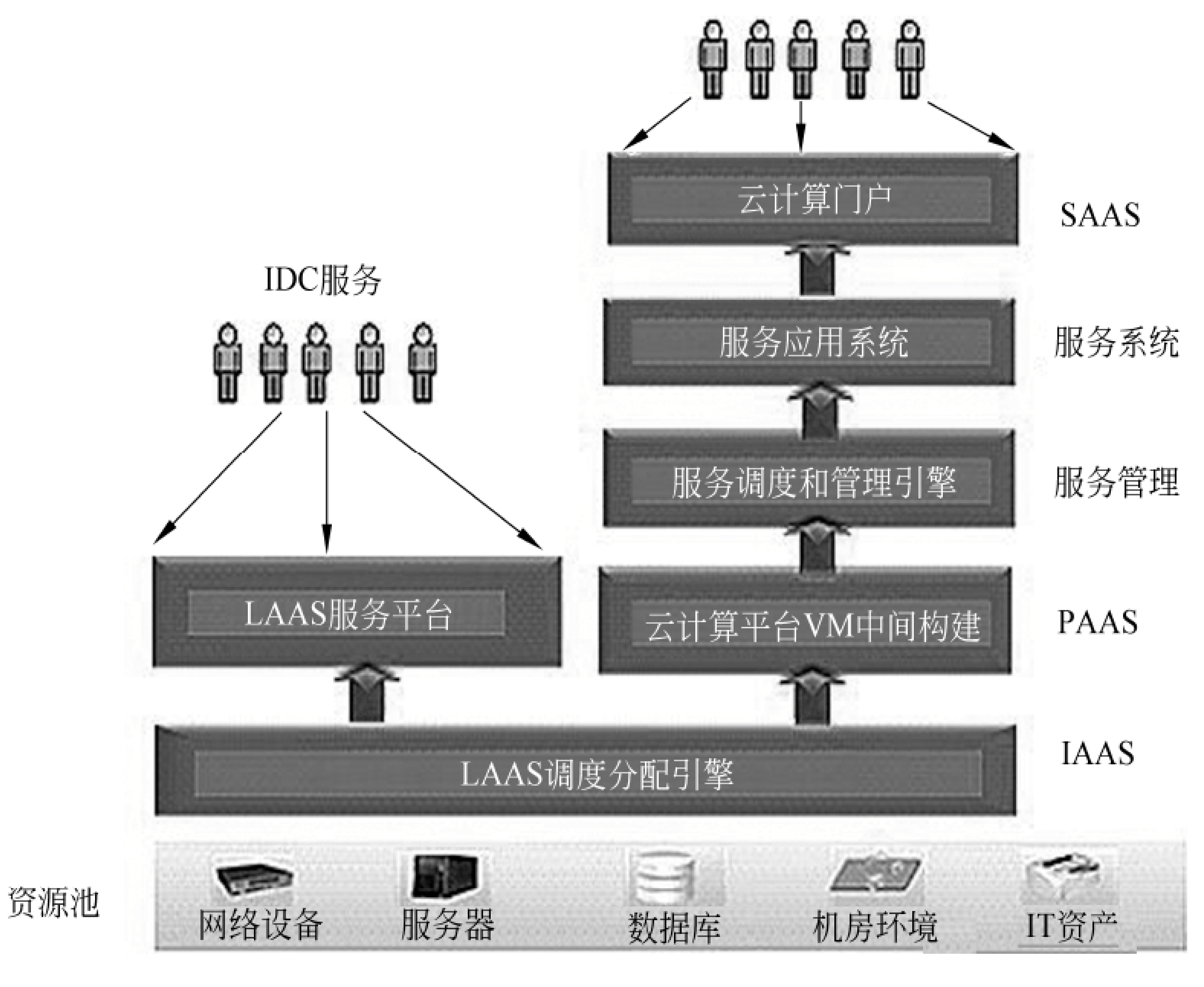


图 2.6　云计算服务层次

2. 云计算产业

云计算的产业三级分层：云软件、云平台、云设备。

1）上层分级：云软件 Software as a Service(SaaS)

打破以往大厂垄断的局面，所有人都可以在上面自由挥洒创意，提供各式各样的软件服务。参与者：世界各地的软件开发者。

2）中层分级：云平台 Platform as a Service(PaaS)

打造程序开发平台与操作系统平台，让开发人员可以通过网络撰写程序与服务，一般消费者也可以在上面运行程序。参与者：Google、微软、苹果、Yahoo!。

3）下层分级：云设备 Infrastructure as a Service(IaaS)

将基础设备(如 IT 系统、数据库等)集成起来，像旅馆一样，分隔成不同的房间供企业租用。参与者：英业达、IBM、戴尔、惠普、亚马逊。

2.2.4　云计算技术层次

云计算技术层次和云计算服务层次不是一个概念，后者从服务的角度来划分云的层次，主要突出了云服务能给用户带来什么。而云计算的技术层次主要从系统属性和设计思想角度来说明云，是对软硬件资源在云计算技术中所充当角色的说明。从云计算技术角度来分，云计算由 4 部分构成：物理资源、虚拟化资源、中间件管理部分和服务接口，如图 2.7 所示。

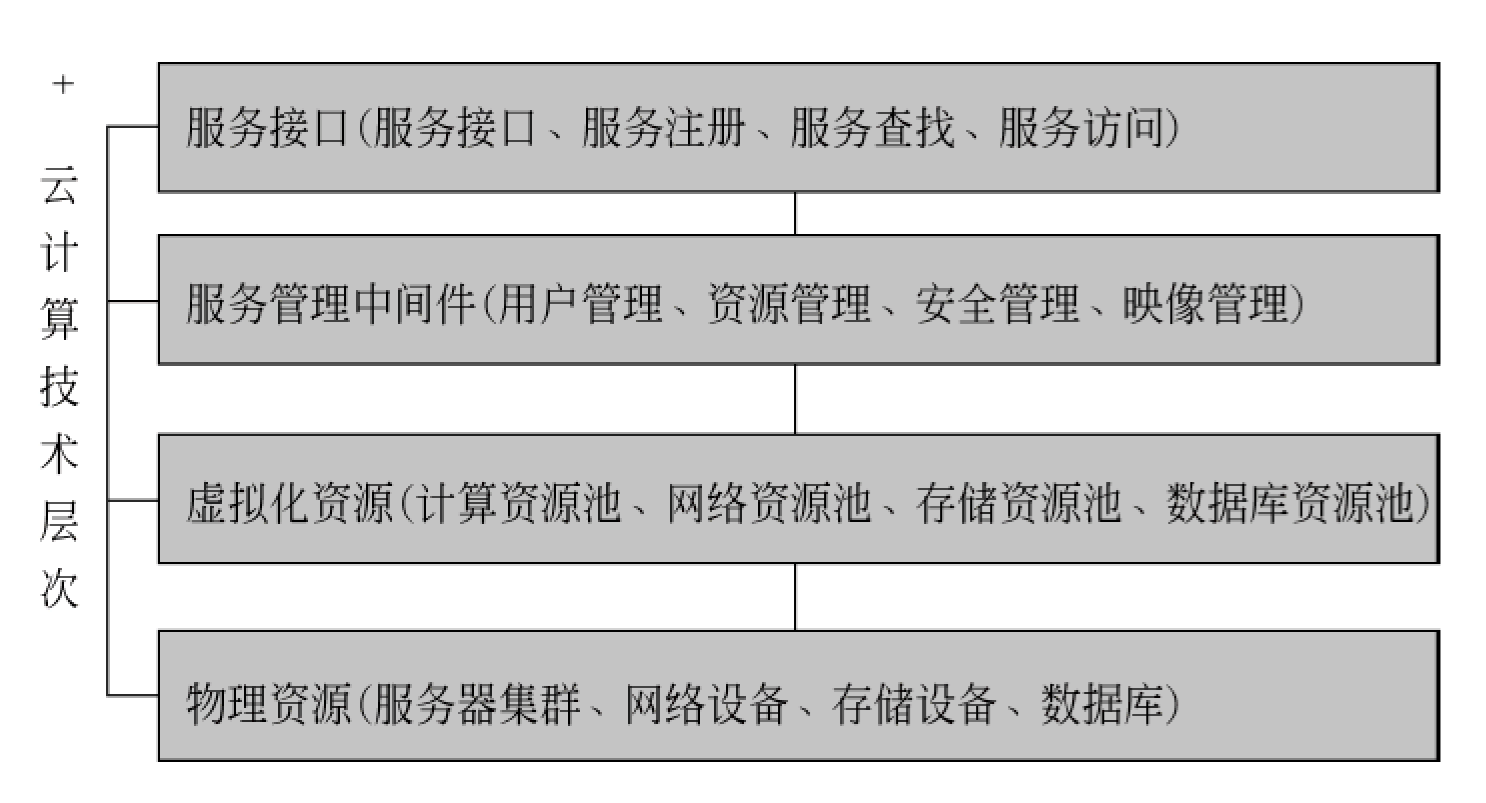


图 2.7　云计算技术层次

1. 服务接口

统一了在云计算时代使用计算机的各种规范、云计算服务的各种标准等，用户端与云端交互操作的入口，可以完成用户或服务注册，对服务的定制和使用。

2. 服务管理中间件

在云计算技术中，中间件位于服务和服务器集群之间，提供管理和服务即云计算体系结构中的管理系统。对标识、认证、授权、目录、安全性等服务进行标准化和操作，为应用提供统一的标准化程序接口和协议，隐藏底层硬件、操作系统和网络的异构性，统一管理网络资源。其用户管理包括用户身份验证、用户许可、用户定制管理；资源管理包括负载

均衡、资源监控、故障检测等;安全管理包括身份验证、访问授权、安全审计、综合防护等;映像管理包括映像创建、部署、管理等。

3. 虚拟化资源

虚拟化资源指一些可以实现一定操作具有一定功能,但其本身是虚拟而不是真实的资源,如计算池,存储池和网络池、数据库资源等,通过软件技术来实现相关的虚拟化功能包括虚拟环境、虚拟系统、虚拟平台。

4. 物理资源

物理资源主要指能支持计算机正常运行的一些硬件设备及技术,可以是价格低廉的PC,也可以是价格昂贵的服务器及磁盘阵列等设备,可以通过现有网络技术和并行技术、分布式技术将分散的计算机组成一个能提供超强功能的集群用于计算和存储等云计算操作。在云计算时代,本地计算机可能不再像传统计算机那样需要空间足够的硬盘、大功率的处理器和大容量的内存,只需要一些必要的硬件设备,如网络设备和基本的输入输出设备等。

2.2.5 云计算的核心技术

云计算系统运用了许多技术,其中以编程模型、数据管理技术、数据存储技术、虚拟化技术、云计算平台管理技术最为关键。

1. 编程模型(MapReduce)

MapReduce 是 Google 开发的 Java、Python、C++ 编程工具,用于大规模数据集(大于1TB)的并行运算,也是云计算的核心技术,一种分布式运算技术,也是简化的分布式编程模式,适合用来处理大量数据的分布式运算,用于解决问题的程序开发模型,也是开发人员拆解问题的方法。

MapReduce 是一种简化的分布式编程模型和高效的任务调度模型,严格的编程模型使云计算环境下的编程十分简单。MapReduce 模式的思想是将要执行的问题分解成 Map(映射)和 Reduce(化简)的方式,先通过 Map 程序将数据切割成不相关的区块,分配(调度)给大量计算机处理,达到分布式运算的效果,再通过 Reduce 程序将结果汇总输出。

2. 海量数据分布存储技术(GFS)

云计算系统由大量服务器组成,同时为大量用户服务,因此云计算系统采用分布式存储的方式存储数据,用冗余存储的方式保证数据的可靠性。云计算系统中广泛使用的数据存储系统是 Google 的 GFS 和 Hadoop 团队开发的 GFS 的开源实现 HDFS。

GFS 即 Google 文件系统(Google File System),是一个可扩展的分布式文件系统,用于大型的、分布式的、对大量数据进行访问的应用。GFS 的设计思想不同于传统的文件系统,是针对大规模数据处理和 Google 应用特性而设计的。它运行于廉价的普通硬件上,但可以提供容错功能。它可以给大量的用户提供总体性能较高的服务。

一个 GFS 集群由一个主服务器(master)和大量的块服务器(chunk server)构成,并

被许多客户(Client)访问。主服务器存储文件系统所有的元数据,包括名字空间、访问控制信息、从文件到块的映射以及块的当前位置。它也控制系统范围的活动,如块租约(lease)管理、孤儿块的垃圾收集、块服务器间的块迁移。

主服务器定期通过 Heart Beat 消息与每一个块服务器通信,给块服务器传递指令并收集它的状态。GFS 中的文件被切分为 64MB 的块并以冗余存储,每份数据在系统中保存 3 个以上备份。

客户与主服务器的交换只限于对元数据的操作,所有数据方面的通信都直接和块服务器联系,这大大提高了系统的效率,防止主服务器负载过重。

3. 海量数据管理技术(BT)

云计算需要对分布的、海量的数据进行处理、分析,因此,数据管理技术必须能够高效地管理大量的数据。云计算系统中的数据管理技术主要是 Google 的 BT(Big Table)数据管理技术和 Hadoop 团队开发的开源数据管理模块 HBase。

BT 是建立在 GFS、Scheduler、Lock Service 和 MapReduce 之上的一个大型的分布式数据库,与传统的关系数据库不同,它把所有数据都作为对象来处理,形成一个巨大的表格,用来分布存储大规模结构化数据。

Google 的很多项目使用 BT 来存储数据,包括网页查询,Google earth 和 Google 金融。这些应用程序对 BT 的要求各不相同:数据大小(从 URL 到网页到卫星图像)不同,反应速度不同(从后端的大批处理到实时数据服务)。对于不同的要求,BT 都成功地提供了灵活高效的服务。

4. 虚拟化技术

通过虚拟化技术可实现软件应用与底层硬件相隔离,它包括将单个资源划分成多个虚拟资源的裂分模式,也包括将多个资源整合成一个虚拟资源的聚合模式。虚拟化技术根据对象可分成存储虚拟化、计算虚拟化、网络虚拟化等,计算虚拟化又分为系统级虚拟化、应用级虚拟化和桌面虚拟化。

5. 云计算平台管理技术

云计算资源规模庞大,服务器数量众多并分布在不同的地点,同时运行着数百种应用,如何有效地管理这些服务器,保证整个系统提供不间断的服务是巨大的挑战。

云计算系统的平台管理技术能够使大量的服务器协同工作,方便地进行业务部署和开通,快速发现和恢复系统故障,通过自动化、智能化的手段实现大规模系统的可靠运营。

2.2.6　典型云计算平台

云计算的研究吸引了不同技术领域巨头,因此对云计算理论及实现架构也有所不同。下面以 Google 公司的云计算核心技术和架构作基本讲解。

云计算的先行者 Google 的云计算平台能实现大规模分布式计算和应用服务程序,平台包括 MapReduce 分布式处理技术、Hadoop 框架、分布式的文件系统 GFS、结构化的 BigTable 存储系统以及 Google 其他的云计算支撑要素。

现有的云计算通过对资源层、平台层和应用层的虚拟化以及物理上的分布式集成,将

庞大的 IT 资源整合在一起。更重要的是，云计算不仅仅是资源的简单汇集，它为我们提供了一种管理机制，让整个体系作为一个虚拟的资源池对外提供服务，并赋予开发者透明获取资源、使用资源的自由。

1．MapReduce 分布式处理技术

MapReduce 是 Google 在 2000 年代初期开发的用于网页索引的用户定义函数。它被设计用来处理分布在多个并行结点的 PB 级和 EB 级数据。

MapReduce 的软件实现是指定一个 Map（映射）函数，把键值对（key/value）映射成新的键值对（key/value），形成一系列中间形式的 key/value 对，然后把它们传给 Reduce（化简）函数，把具有相同中间形式 key 的 value 合并在一起。Map 和 Reduce 函数具有一定的关联性。可以进行海量数据分割、任务分解与结果汇总，从而完成海量数据的并行处理，如图 2.8 所示。

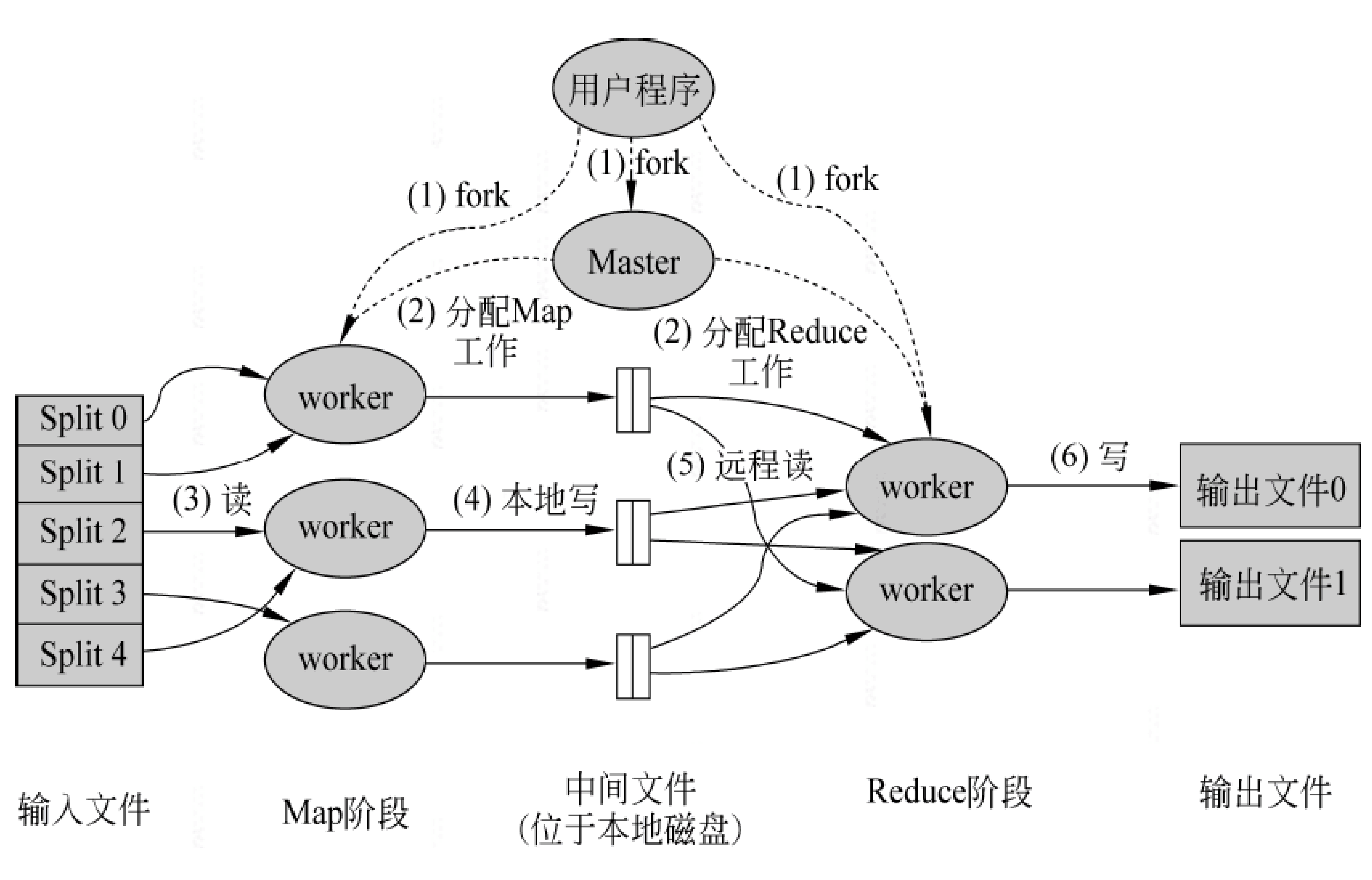


图 2.8　大数据的并行处理利器——MapReduce

2．MapReduce 架构设计

MapReduce 基础出发点是易懂。它由称为 Map 和 Reduce 的两部分用户程序组成，然后利用框架在计算机集群上面根据需求运行多个程序实例来处理各个子任务，然后再对结果进行归并，如图 2.9 所示。

MapReduce 的工作原理其实是先分后合的数据处理方式。Map 即“分解”，把海量数据分割成了若干部分，分给多台处理器并行处理；Reduce 即“合并”，把各台处理器处理后的结果进行汇总操作以得到最终结果。如果采用 MapReduce 来统计不同几何形状的数量，它会先把任务分配到两个结点，由两个结点分别并行统计，然后再把它们的结果汇总，得到最终的计算结果。MapReduce 执行流程如图 2.10 所示。

3．Hadoop 架构

Hadoop 是一个处理、存储和分析海量的分布式、非结构化数据的开源框架。最初由雅虎的 Doug Cutting 创建，Hadoop 的灵感来自于 MapReduce，Hadoop 集群运行在廉价

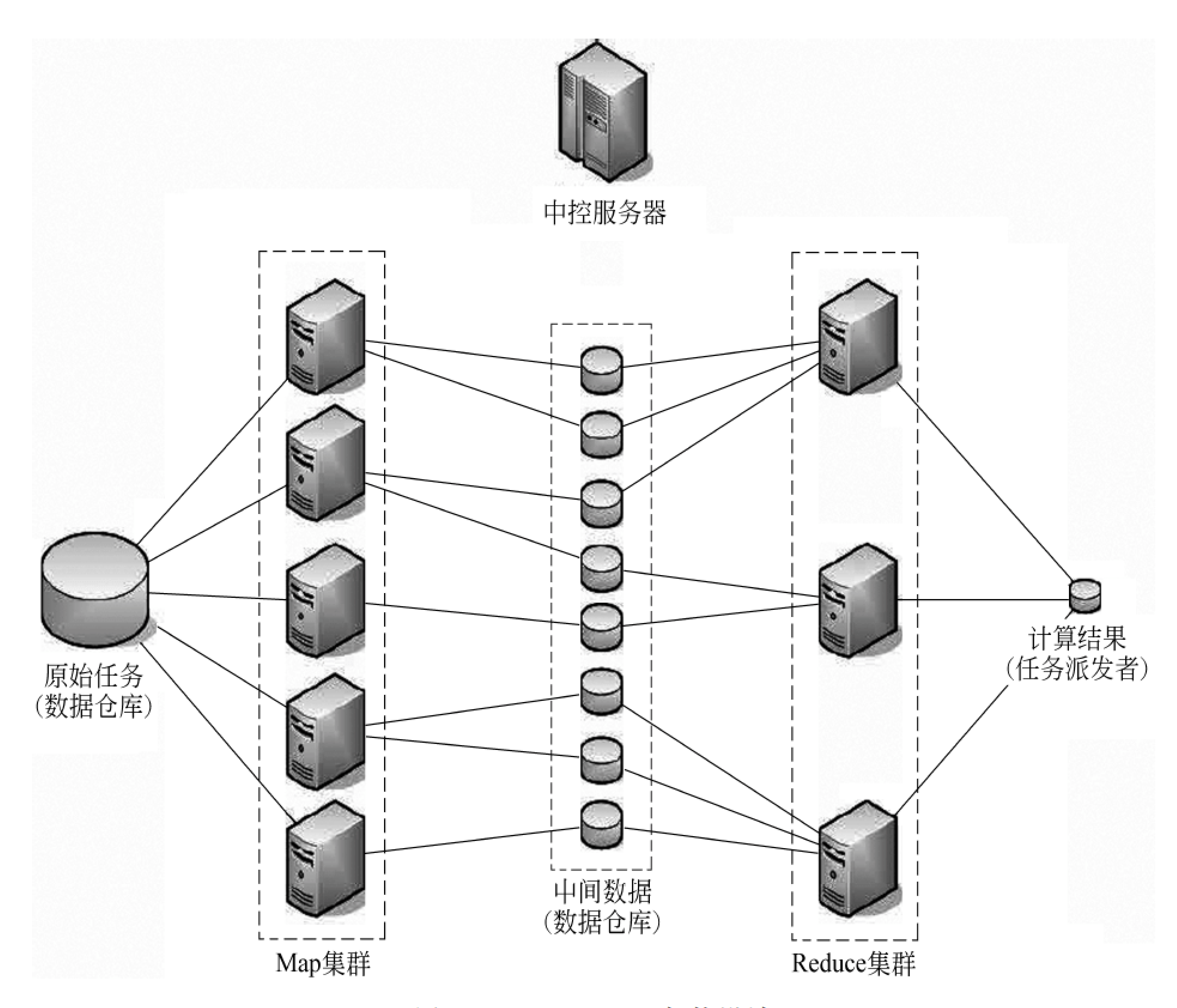


图 2.9　MapReduce 架构设计

MapReduce执行流程
input
File
Split
part1
part2
part3
part4
part5
Map
Map
Map
Map
Map
Suffle
Reduce
Reduce
output
output

图 2.10　MapReduce 执行流程

的商用硬件上，这样硬件扩展就不存在资金压力。Hadoop 现在是 Apache 软件联盟(The Apache Software Foundation)的一个项目，数百名贡献者不断改进其核心技术。

其基本概念与将海量数据限定在一台机器运行的方式不同，Hadoop 将大数据分成多个部分，这样每个部分都可以被同时处理和分析。

在 Google 发表 MapReduce 后，2004 年开源社群用 Java 搭建出一套 Hadoop 框架，

用于实现 MapReduce 算法,能够把应用程序分割成许多很小的工作单元,每个单元可以在任何集群结点上执行或重复执行。

此外,Hadoop 还提供一个分布式文件系统 GFS(Google File System),是一个可扩展、结构化、具备日志的分布式文件系统,支持大型、分布式大数据量的读写操作,其容错性较强。

而分布式数据库(BigTable)是一个有序、稀疏、多维度的映射表,有良好的伸缩性和高可用性,用来将数据存储或部署到各个计算结点上。Hadoop 框架具有高容错性及对数据读写的高吞吐率,能自动处理失败结点,如图 2.11 所示为 Google Hadoop 架构。

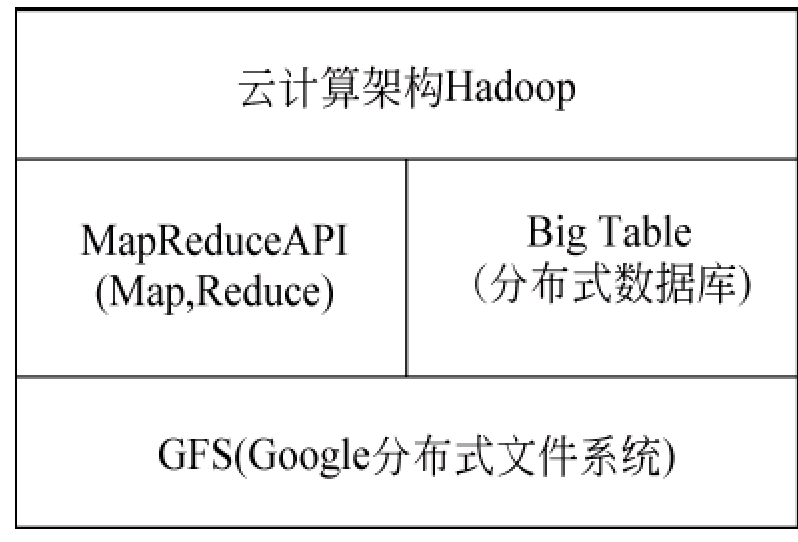


图 2.11　Hadoop 架构

在架构中 MapReduce API 提供 Map 和 Reduce 处理、GFS 分布式文件系统和 BigTable 分布式数据库提供数据存取。基于 Hadoop 可以非常轻松和方便地完成处理海量数据的分布式并行程序,并运行于大规模集群上。

1）Hadoop 如何工作

客户从日志文件、社交媒体供稿和内部数据存储等来源获得非结构化和半结构化数据。它将数据打碎成“部分”,这些“部分”被载入到商用硬件的多个结点组成的文件系统。Hadoop 的默认文件存储系统是 Hadoop 分布式文件系统。文件系统(如 HDFS)适合存储大量非结构化和半结构化数据,因为它们不需要将数据组织成关系型的行和列。

各“部分”被复制多次,并加载到文件系统。这样,如果一个结点失效,另一个结点包含失效结点数据的副本。名称结点充当调解人,负责沟通信息:如哪些结点是可用的,某些数据存储在集群的什么地方,以及哪些结点失效。

一旦数据被加载到集群中,它就准备好通过 MapReduce 框架进行分析。客户提交一个“匹配”的任务(通常是用 Java 编写的查询语句)给到一个被称为作业跟踪器的结点。该作业跟踪器引用名称结点,以确定完成工作需要访问哪些数据,以及所需的数据在集群的存储位置。一旦确定,作业跟踪器向相关结点提交查询。每个结点同时、并行处理,而非将所有数据集中到一个位置处理。这是 Hadoop 的一个本质特征。

当每个结点处理完指定的作业,它会存储结果。客户通过任务追踪器启动 Reduce 任务。汇总 Map 阶段存储在各个结点上的结果数据,获得原始查询的“答案”,然后将“答案”加载到集群的另一个结点中。客户就可以访问这些可以载入多种分析环境进行分析的结果了。MapReduce 的工作就完成了。

一旦 MapReduce 阶段完成,数据科学家和其他人就可以使用高级数据分析技巧对处理后的数据进一步分析。也可以对这些数据建模,将数据从 Hadoop 集群转移到现有的关系型数据库、数据仓库等传统 IT 系统进行进一步的分析。

Hadoop 的三大核心设计如图 2.12 所示。

2）Hadoop 的技术组件

Hadoop“栈”由多个组件组成。包括:

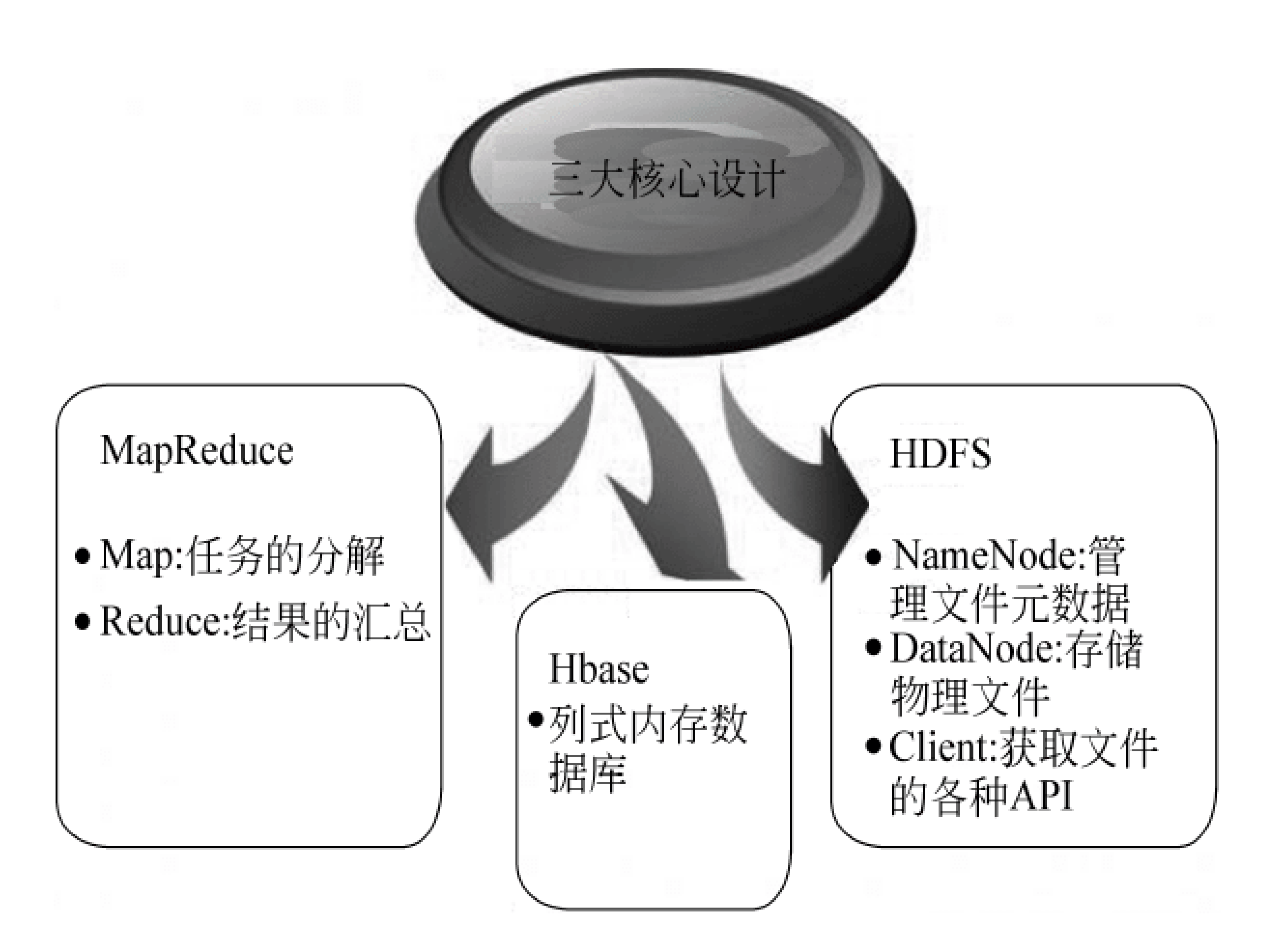


图 2.12　Hadoop 三大核心设计

- Hadoop 分布式文件系统(HDFS)——所有 Hadoop 集群的默认存储层;
- 名称结点——在 Hadoop 集群中,提供数据存储位置以及结点失效信息的结点。
- 二级结点——名称结点的备份,它会定期复制和存储名称结点的数据,以防名称结点失效。
- 作业跟踪器——Hadoop 集群中发起和协调 MapReduce 作业或数据处理任务的结点。
- 从结点——Hadoop 集群的普通结点,从结点存储数据并且从作业跟踪器那里获取数据处理指令。

除了上述内容以外,Hadoop 生态系统还包括许多免费子项目。NoSQL 数据存储系统(如 Cassandra 和 HBase)也被用于存储 Hadoop 的 MapReduce 作业结果。除了 Java,很多 MapReduce 作业及其他 Hadoop 的功能都是用 Pig 语言写的,Pig 是专门针对 Hadoop 设计的开源语言。Hive 最初是由 Facebook 开发的开源数据仓库,可以在 Hadoop 中建立分析模型。

3) Hadoop:优点和缺点

Hadoop 的主要好处是,它可以让企业以节省成本并以高效的方式处理和分析大量的非结构化和半结构化数据,而这类数据迄今还没有其他处理方式。因为 Hadoop 集群可以扩展到 PB 级甚至 EB 级数据,企业不再必须依赖于样本数据集,而可以处理和分析所有相关数据。数据科学家可以采用迭代的方法进行分析,不断改进和测试查询语句,从而发现以前未知的见解。使用 Hadoop 的成本也很廉价。开发者可以免费下载 Apache 的 Hadoop 分布式平台,并且在不到一天的时间内开始体验 Hadoop。

Hadoop 及其无数组件的不足之处是,它们还不成熟,仍处于发展阶段。就像所有新的、原始的技术一样,实施和管理 Hadoop 集群,对大量非结构化数据进行高级分析,都需要大量的专业知识、技能和培训。不幸的是,目前 Hadoop 开发者和数据科学家的缺乏,使得众多企业维持复杂的 Hadoop 集群并利用其优势变得很不现实。

此外,由于 Hadoop 的众多组件都是通过技术社区得到改善,并且新的组件不断被创建,因此作为不成熟的开源技术,也存在失败的风险。最后,Hadoop 是一个面向批处理

的框架，这意味着它不支持实时的数据处理和分析。

4. Google 云计算执行过程

云计算服务方式多种多样，通过对 Google 云计算架构及技术的理解，在此我们给出用户将要执行的程序或处理的问题提交云计算的平台 Hadoop，其执行过程如图 2.13 所示。

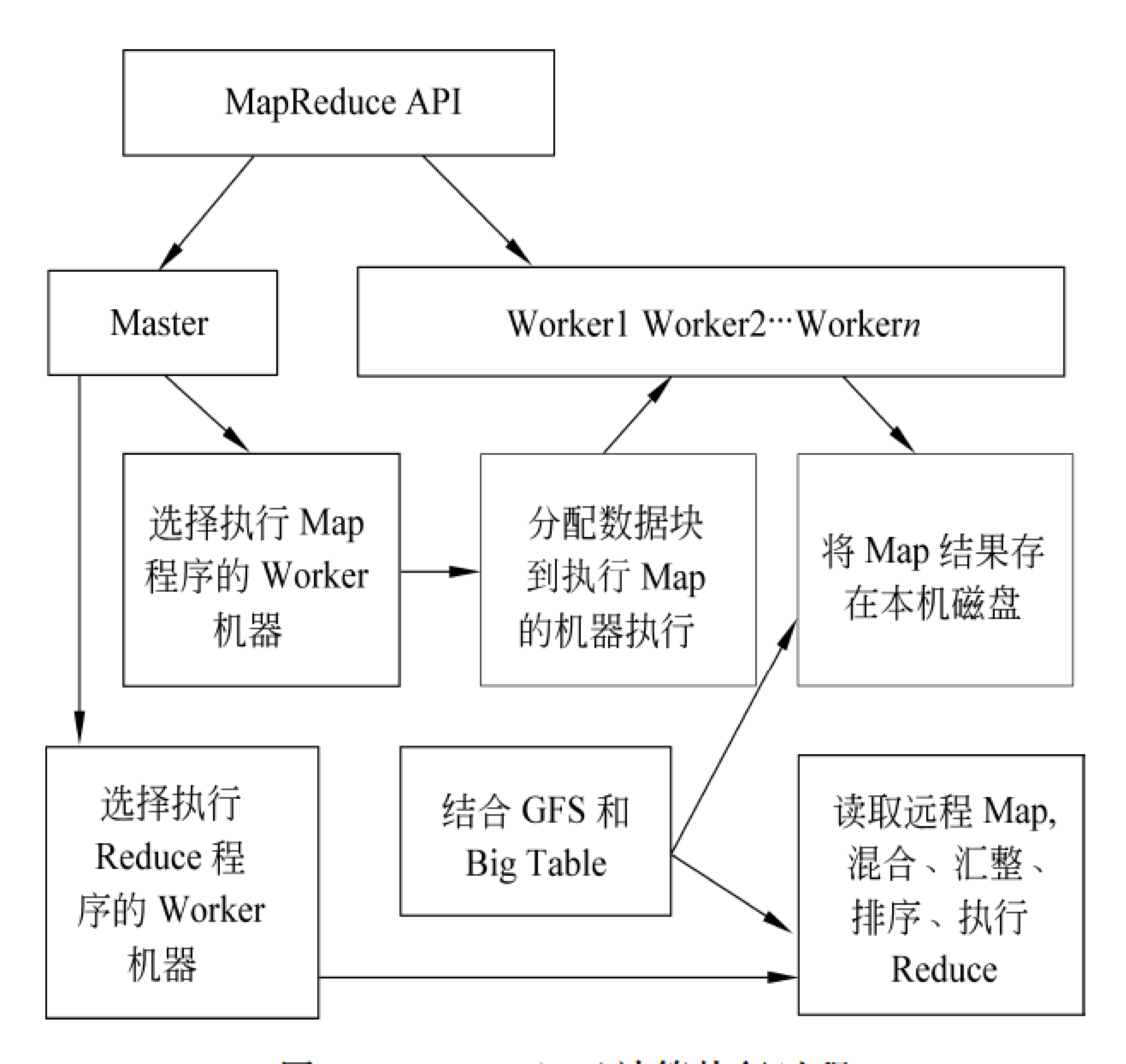


图 2.13　Google 云计算执行过程

如图 2.13 所示的 Google 云计算执行过程包括以下步骤：

(1) 将要执行的 MPI 程序复制到 Hadoop 框架中的 Master 和每一台 Worker 机器中。

(2) Master 选择由哪些 Worker 机器来执行 Map 程序与 Reduce 程序。

(3) 分配所有的数据区块到执行 Map 程序的 Worker 机器中进行 Map(切割成小块数据)。

(4) 将 Map 后的结果存入 Worker 机器。

(5) 执行 Reduce 程序的 Worker 机器，远程读取每一份 Map 结果，进行混合、汇整与排序，同时执行 Reduce 程序。

(6) 将结果输出给用户(开发者)。

在云计算中为了保证计算和存储等操作的完整性，充分利用 MapReduce 的分布和可靠特性，在数据上传和下载过程中根据各 Worker 结点在指定时间内反馈的信息判断结点的状态是正常还是死亡。若结点死亡，则将其负责的任务分配给别的结点，以确保文件数据的完整性。

2.2.7　典型的云计算系统及应用

由于云计算技术范围很广，目前各大 IT 企业提供的云计算服务主要根据自身的特

点和优势实现的。下面以 Google、IBM、Amazon 为例说明。

1. Google 的云计算平台

Google 的硬件条件优势，大型的数据中心、搜索引擎的支柱应用，促进 Google 云计算迅速发展。Google 的云计算主要由 MapReduce、Google 文件系统（GFS）、BigTable 组成。它们是 Google 内部云计算基础平台的 3 个主要部分。Google 还构建其他云计算组件，包括一个领域描述语言以及分布式锁服务机制等。Sawzall 是一种建立在 MapReduce 基础上的领域语言，专门用于大规模的信息处理。Chubby 是一个高可用、分布式数据锁服务，当有机器失效时，Chubby 使用 Paxos 算法来保证备份。

2. IBM“蓝云”计算平台

“蓝云”解决方案是由 IBM 云计算中心开发的企业级云计算解决方案。“蓝云”基于 IBM Almaden 研究中心的云基础架构，采用了 Xen 和 PowerVM 虚拟化软件，Linux 操作系统映像以及 Hadoop 软件（Google File System 以及 MapReduce 的开源实现）。

“蓝云”计算平台由一个数据中心、IBM Tivoli 部署管理软件（Tivoli provisioning manager）、IBM Tivoli 监控软件（IBM Tivoli monitoring）、IBM WebSphere 应用服务器、IBM DB2 数据库以及一些开源信息处理软件和开源虚拟化软件共同组成。“蓝云”的硬件平台环境与一般的 x86 服务器集群类似，使用刀片的方式增加了计算密度。“蓝云”软件平台的特点主要体现在虚拟机以及对于大规模数据处理软件 Apache Hadoop 的使用上。

“蓝云”平台的一个重要特点是虚拟化技术的使用。虚拟化的方式在“蓝云”中有两个级别：一个是在硬件级别上实现虚拟化，另一个是通过开源软件实现虚拟化。硬件级别的虚拟化可以使用 IBM p 系列的服务器，获得硬件的逻辑分区 LPAR（logic partition）。逻辑分区的 CPU 资源能够通过 IBM Enterprise Workload Manager 来管理。通过这样的方式加上在实际使用过程中的资源分配策略，能够使相应的资源合理地分配到各个逻辑分区。p 系列系统的逻辑分区最小粒度是 1/10 颗 CPU。Xen 则是软件级别上的虚拟化，能够在 Linux 基础上运行另外一个操作系统。

“蓝云”存储体系结构包含类似于 Google File System 的集群文件系统以及基于块设备方式的存储区域网络 SAN。在设计云计算平台的存储体系结构时，可以通过组合多个磁盘获得很大的磁盘容量。相对于磁盘的容量，在云计算平台的存储中，磁盘数据的读写速度是一个更重要的问题，因此需要对多个磁盘进行同时读写。这种方式要求将数据分配到多个结点的多个磁盘当中。为达到这一目的，存储技术有两个选择：一个是使用类似于 Google File System 的集群文件系统，另一个是基于块设备的存储区域网络 SAN 系统。

3. Amazon 的弹性计算云

Amazon 是互联网上最大的在线零售商，为了应付交易高峰，不得不购买了大量的服务器。而在大多数时间，大部分服务器闲置，造成了很大的浪费，为了合理利用空闲服务器，Amazon 建立了自己的云计算平台弹性计算云 EC2（Elastic Compute Cloud），并且是第一家将基础设施作为服务出售的公司。

Amazon 将自己的弹性计算云建立在公司内部的大规模集群计算的平台上，而用户可以通过弹性计算云的网络界面去操作在云计算平台上运行的各个实例(instance)。用户使用实例的付费方式由用户的使用状况决定，即用户只需为自己所使用的计算平台实例付费，运行结束后计费也随之结束。这里所说的实例即是由用户控制的完整的虚拟机运行实例。通过这种方式，用户不必自己去建立云计算平台，节省了设备与维护费用。

弹性计算云用户使用客户端通过 SOAP over HTTPS 协议与 Amazon 弹性计算云内部的实例进行交互。这样，弹性计算云平台为用户或者开发人员提供了一个虚拟的集群环境，在用户具有充分灵活性的同时，也减轻了云计算平台拥有者(Amazon 公司)的管理负担。弹性计算云中的每一个实例代表一个运行中的虚拟机。用户对自己的虚拟机具有完整的访问权限，包括针对此虚拟机操作系统的管理员权限。虚拟机的收费也是根据虚拟机的能力进行费用计算的，实际上，用户租用的是虚拟的计算能力。

总而言之，Amazon 通过提供弹性计算云，满足了小规模软件开发人员对集群系统的需求，减小了维护负担。其收费方式相对简单明了：用户使用多少资源，只需为这一部分资源付费即可。

为了弹性计算云的进一步发展，Amazon 规划了如何在云计算平台基础上帮助用户开发网络化的应用程序。除了网络零售业务以外，云计算也是 Amazon 公司的核心价值所在。Amazon 将来会在弹性计算云的平台基础上添加更多的网络服务组件模块，为用户构建云计算应用提供方便。

4. 云计算系统间的特性比较

从用户的角度来看，云计算系统将各种数据包括用户数据都通过网络保存到远端的云存储平台上，减小了用户对于数据管理的负担；同时，云计算系统也将处理数据的服务程序通 过远程的大规模云计算处理平台进行，能够负担大量数据的处理工作。可以说，云计算是数据共享计算模式与服务共享计算模式的结合体，是下一代计算模式的发展方向。

各个云计算平台各自具有不同的特点。特别是在平台的使用上，透明计算平台为用户同时提供了用户实际接触的客户端结点以及无法接触的远程虚拟存储服务器，是一个半公开的环境。表 2.1 从多个角度比较了各个云计算系统的不同之处。可以看出，虽然云计算系统在很多方面具有共性，但实际上各个系统之间还是有很大不同的，这也给云计算用户或者开发人员带来了不同的体验。

表 2.1 各个云计算系统的比较

云计算平台特性	Google 云计算架构	IBM 云计算产品	亚马逊弹性计算云
与传统软件的兼容性	在搜索基础上建立的新的网络系统；当前的软件还不能在该架构下运行，无兼容性	采用了虚拟技术，既能运行传统软件又能提供新的云计算接口给新应用程序开发	采用了虚拟技术，可以运行传统软件
系统的开放性	采用内部技术	采用开源技术	结合内部技术和开源技术
系统虚拟技术的采用	未采用系统虚拟技术，只能支持新应用	采用开源虚拟软件 Xen	采用开源虚拟软件 Xen

续表

云计算平台特性	Google 云计算架构	IBM 云计算产品	亚马逊弹性计算云
目标用户	用户可以直接使用,同时提供网络应用程序编程标准给开发人员	开发人员	开发人员
编程支持	提供网络应用程序编程标准	局部分布式应用程序编程接口	网络远程操作接口

2.2.8　大数据平台的应用

1. 传统处理平台已不适应大数据的处理

大数据环境下数据来源非常丰富且数据类型多样,存储和分析挖掘的数据量庞大,对数据展现的要求较高,并且很看重数据处理的高效性和可用性。

传统的数据采集来源单一,且存储、管理和分析数据量也相对较小,大多采用关系型数据库和并行数据仓库即可处理。对依靠并行计算提升数据处理速度方面而言,传统的并行数据库技术追求高度一致性和容错性,根据 CAP 理论,难以保证其可用性和扩展性。

传统的数据处理方法是以处理器为中心,而在大数据环境下,需要采取以数据为中心的模式,减少数据移动带来的开销。因此,传统的数据处理方法,已经不能适应大数据的需求!

2. 大数据平台的处理方式

大数据的基本处理流程与传统数据处理流程并无太大差异,主要区别在于:由于大数据要处理大量、非结构化的数据,所以在各个处理环节中都可以采用 MapReduce 等方式进行并行处理,如图 2.14 所示。

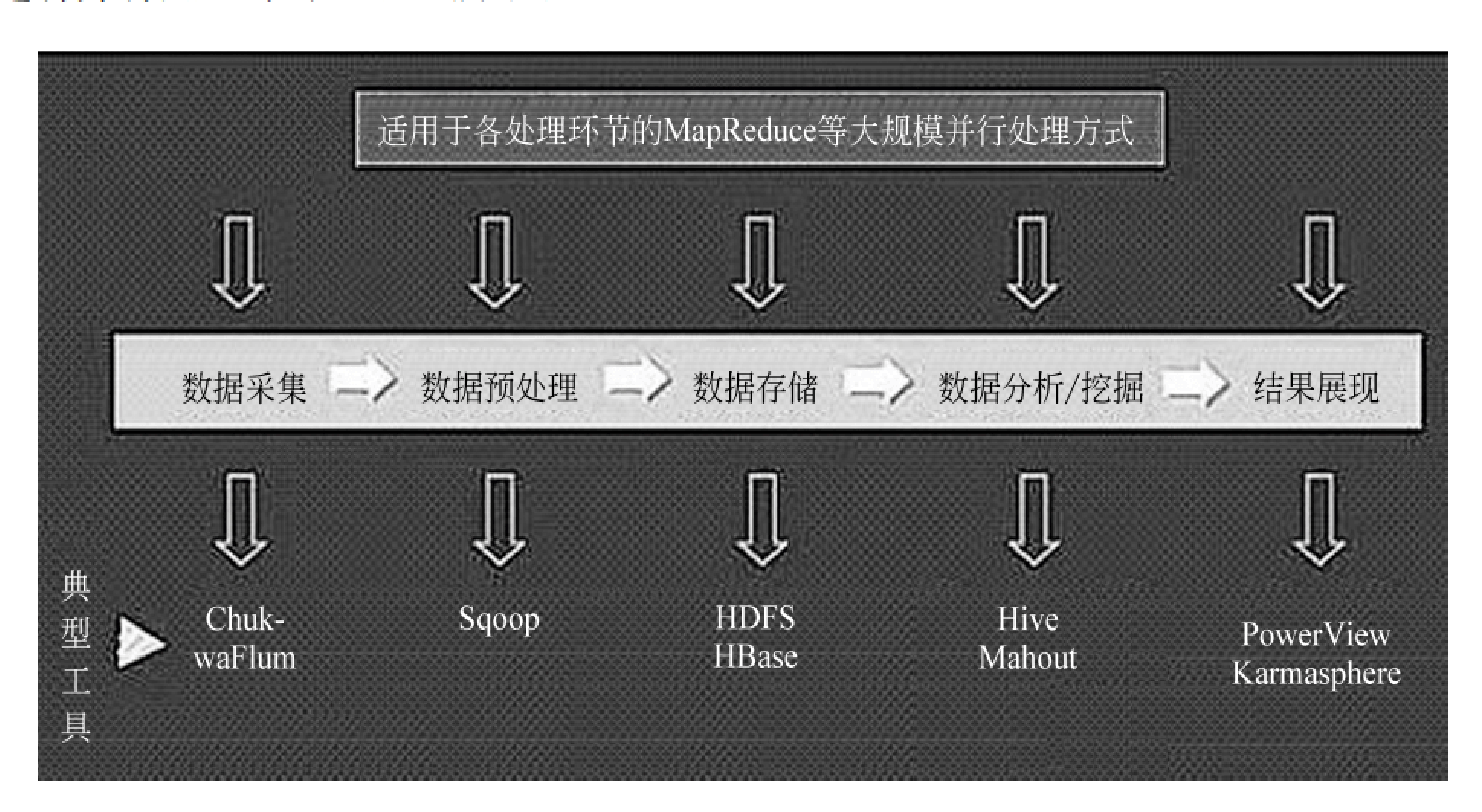


图 2.14　大数据平台的处理方式

3. 大数据技术为什么能提高数据的处理速度

大数据可以通过 MapReduce 这一并行处理技术来提高数据的处理速度。

MapReduce 的设计初衷是通过大量廉价服务器实现大数据并行处理，对数据一致性要求不高，其突出优势是具有扩展性和可用性，特别适用于海量的结构化、半结构化及非结构化数据的混合处理，如图 2.15 所示。

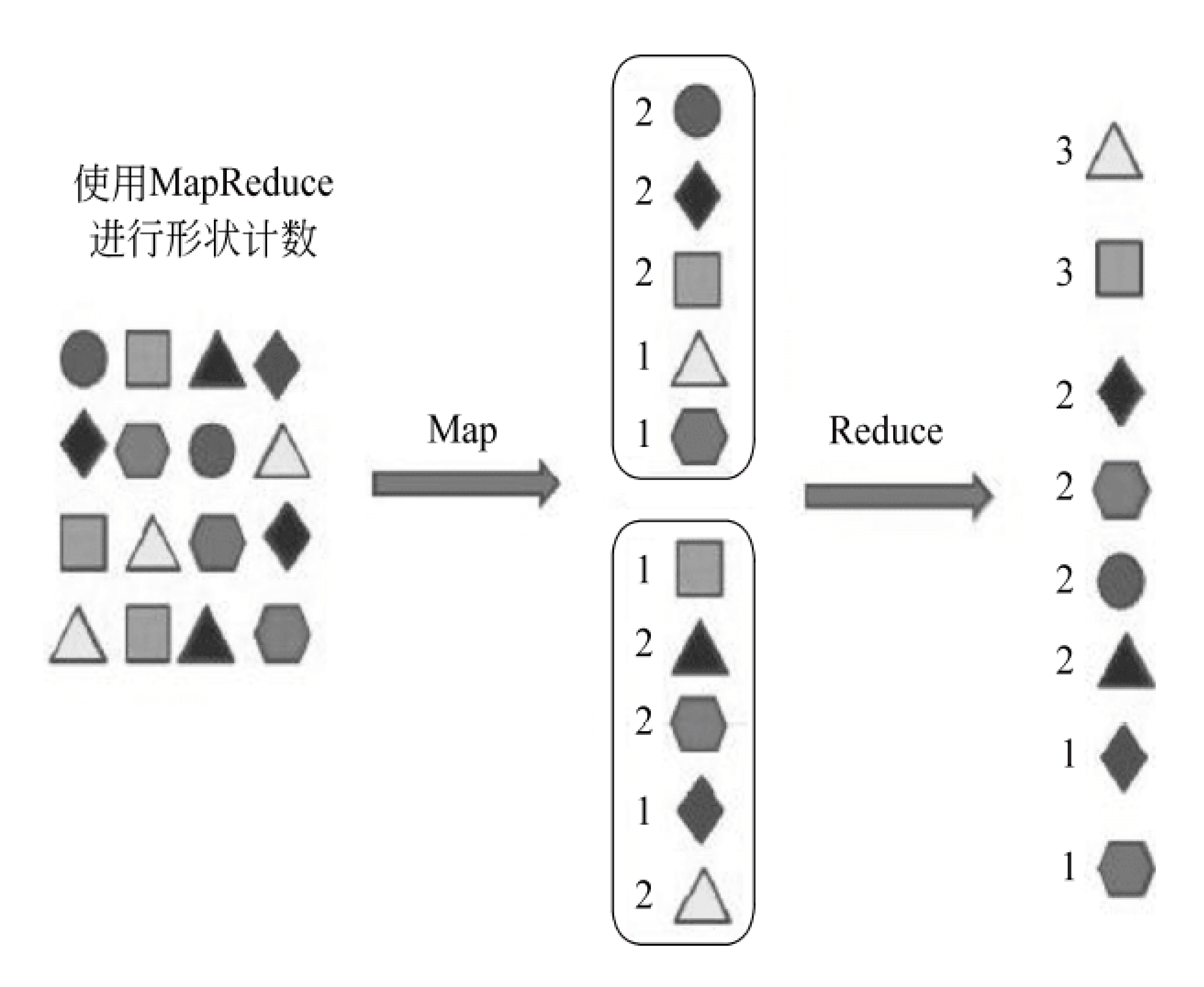


图 2.15　MapReduce 技术进行实时分析

MapReduce 将传统的查询、分解及数据分析进行分布式处理，将处理任务分配到不同的处理结点，因此具有更强的并行处理能力。作为一个简化的并行处理的编程模型，MapReduce 还降低了开发并行应用的门槛。

MapReduce 适合进行数据分析、日志分析、商业智能分析、客户营销、大规模索引等业务，并具有非常明显的效果。通过结合 MapReduce 技术进行实时分析，某家电公司的信用计算时间从 33 小时缩短到 8 秒，而 MKI 的基因分析时间从数天缩短到 20 分钟。

说到这里，再看一看 MapReduce 与传统的分布式并行计算环境 MPI 到底有何不同？MapReduce 在其设计目的、使用方式以及对文件系统的支持等方面与 MPI 都有很大的差异，使其能够更加适应大数据环境下的处理需求，如表 2.2 所示。

表 2.2　MapReduce 与传统的分布式并行计算环境 MPI 的区别

	MapReduce	MPI
设计目的	用于互联网服务 使用大量廉价 PC 耦合度低 结点失效率高 有容错机制	用于科学计算 多使用专用并行机 耦合度高 结点失效率低 无备份
使用方式	以架构形式提出 系统自动选择计算结点，分布处理对用户透明	提供结点间信息沟通的工具，架构不固定 计算结点由开发者指定
对文件系统的支持	支持分布式文件系统 通过 MapReduce 函数实现分布并行计算	不支持分布式文件系统，数据集中存储 由高级语言通过调用标准函数传递消息实现并行计算

2.3　大数据应用案例之：在“北上广”打拼是怎样一种体验

到“北上广”等大都市去闯荡、打拼，是很多年轻人的梦想。即便是在高房价、高物价、交通拥堵、空气污染下被迫离开的人，也有相当一部分重新回来。这些远离亲人，选择面对生活的艰苦和孤独的年轻人，究竟是怎样的群体，又过着什么样的生活？通过大数据分析，你或许能了解一二。

1. 北上广的“飘”们都来自哪里

根据卫计委 2014 年数据，全国 9433 万跨省流动人口，超过 1/5 涌入了北京、上海、广州三个城市。特别是广州，外来人口数量已经超过了常住户籍人口，而在北京和上海，本地人和外地人的比例分别是 1.6∶1 和 1.44∶1，如图 2.16 所示。

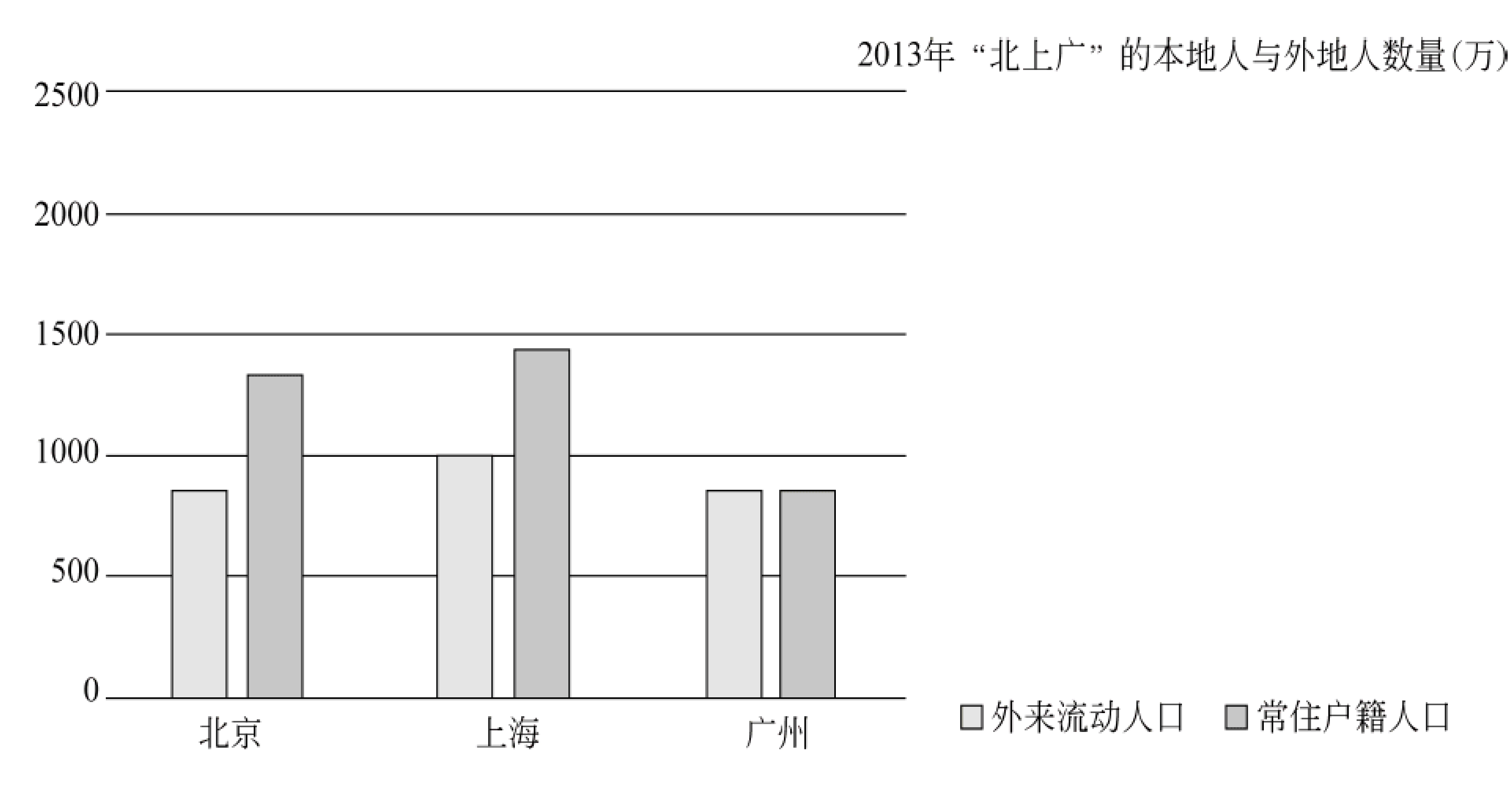


图 2.16　2013 年“北上广”的本地人与外地人数量(万)

从外来人口来源省份看，北京、上海、广州分别在华北、华中、华南地区以吸收周边邻省人口为主。而作为人口流出大省的河南、湖北，则同时进入了“北上广”外来人口数量排名的前五，可见其南北通吃、势力强大。

2. 年纪轻、学历高，或更能站稳脚跟

在“北上广”，拼搏奋斗的核心人群在 20～40 岁之间，占整体外来人口比例都超过 75%。但从年龄结构比较，上海的年轻群体年龄段更为集中，北京 45 岁以上人群占比明显大于其他，而广州外来人口的年龄构成则更偏向年轻化。

2012 年，国家人口计生委曾对“北上广”35 岁以下青年流动人口的生活状态作过监测研究。发现收入是影响其生活质量的重要因素之一，更是坚守或逃离“北上广”的关键。

影响收入最关键的因素被认为是学历。“北上广”三地学历在本科以上的外来青年，月均收入分别是 5652 元、5756 元和 6569 元，详见图 2.17。

“流动中国”调查数据显示，广州本科及以上学历的青年人群比例确实远低于北京和上海，这或许是高学历年轻人在广州更“吃香”的一个原因。

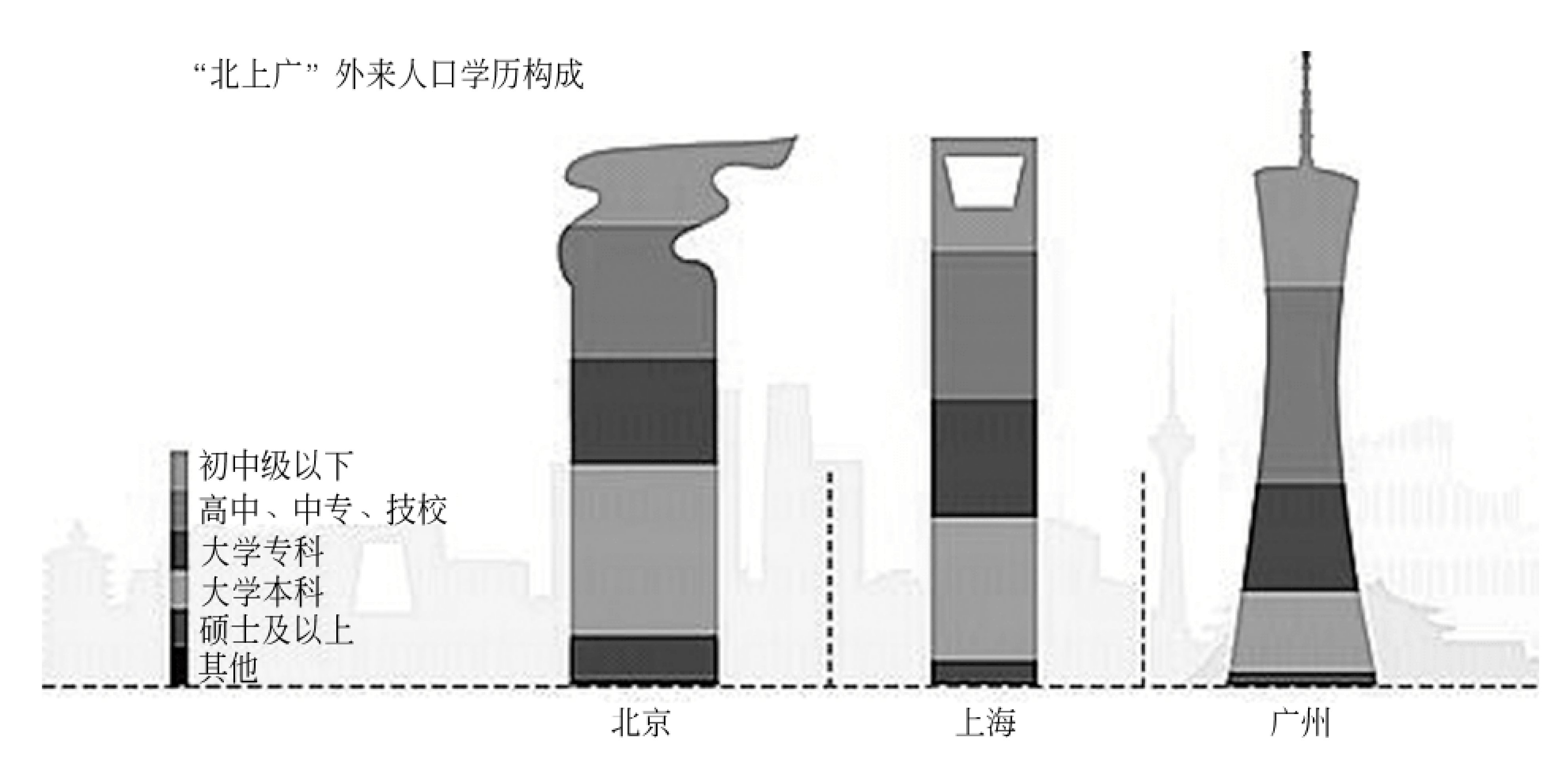


图 2.17 外来人口学历构成

另外，在上海、广州的外来年轻人和全国同龄流动人口一样，以从事制造业为主，约占四成左右，其次是批发零售、建筑、社会服务等行业。

不过，北京的情况较为不同，从事制造业的比重明显较低，从事互联网、金融、房地产的明显高于其他二者。这与北京外来青年学历层次较高及城市功能定位有关，详见图 2.18。

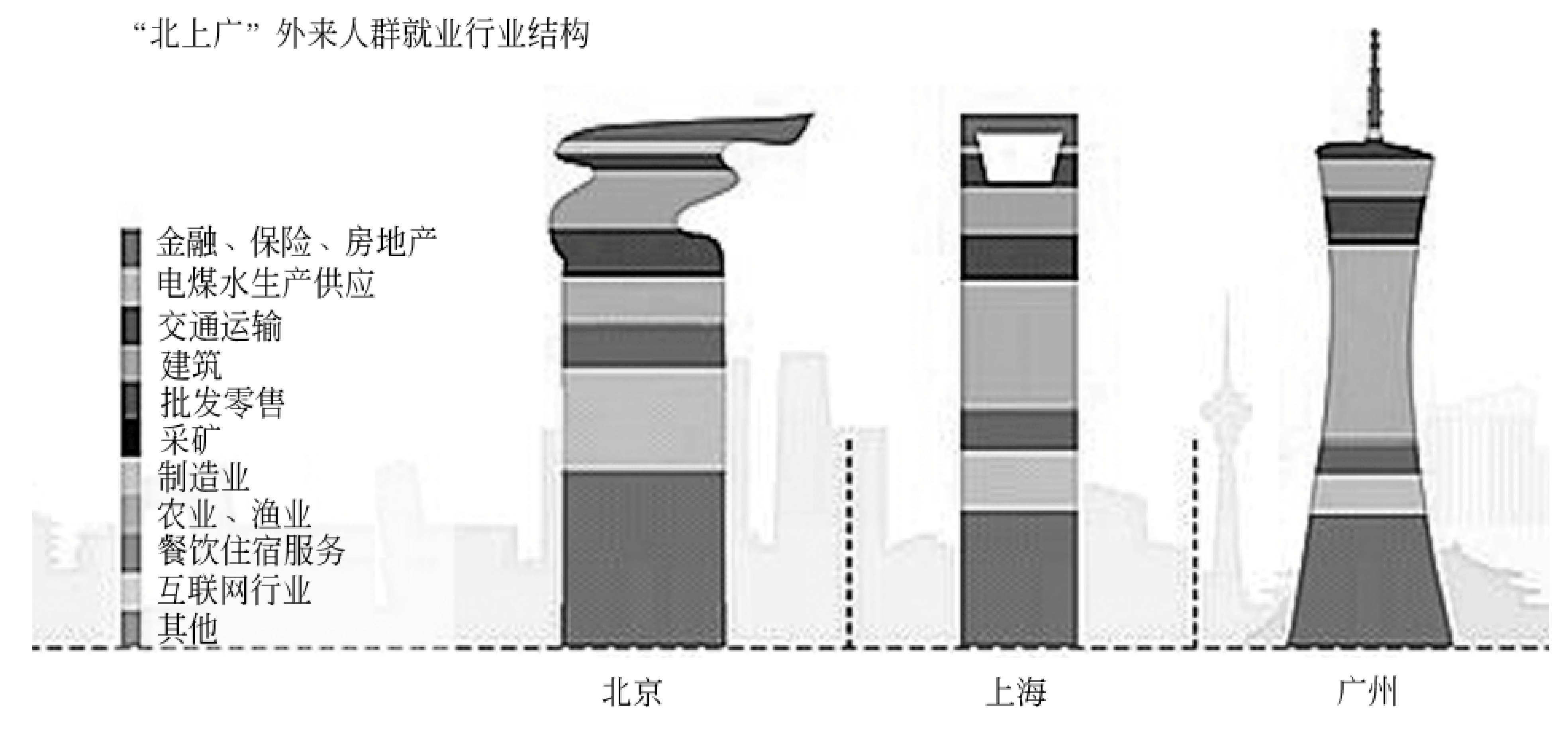


图 2.18 外来人口就业行业构成

3. 一样的"飘"，却分出了上、中、下

在"北上广"三地，外来人口的住房情况大体一致，均有过半数人租房居住。北京人均租房平均月支出 904 元，超过全国平均水平 70%，几乎是食品月支出的两倍。可见租房的花销最让"北漂"们肉痛。"流动中国"调查数据中，广州的老板们能给解决住宿的比例最高，这一点格外明显，详见图 2.19。

当然，在不同历史和政策背景下，"北上广"三地也均形成了外来人口聚居的城中村，

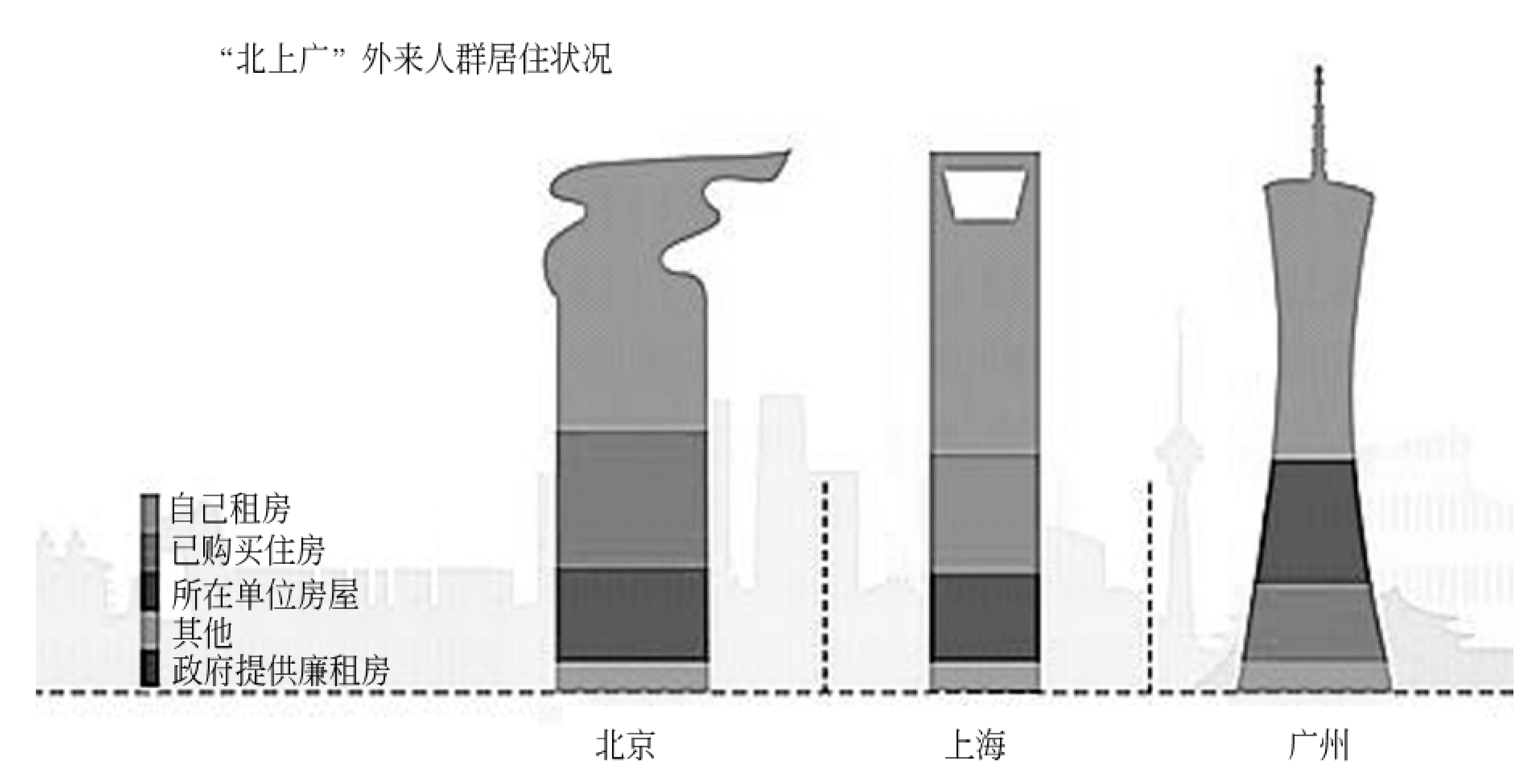


图 2.19　外来人群居住状态

作为多数人"停泊"的首站。随着房价持续上涨，北京的"蚁族"、上海的"蜗居"一族曾一度在公众中流行。

比较"北上广"的城中村，着实是一个有趣的话题，如图 2.20 所示的外来人群居住状态及房屋空间变化呈现了其中的不同。广州的城中村散布在城市中的各个角落，规模和占地都较大；上海的则分布在内环外靠近外围地区，且规模较小；北京城中村主要分布在城市建成区边缘地带，约为五环附近。

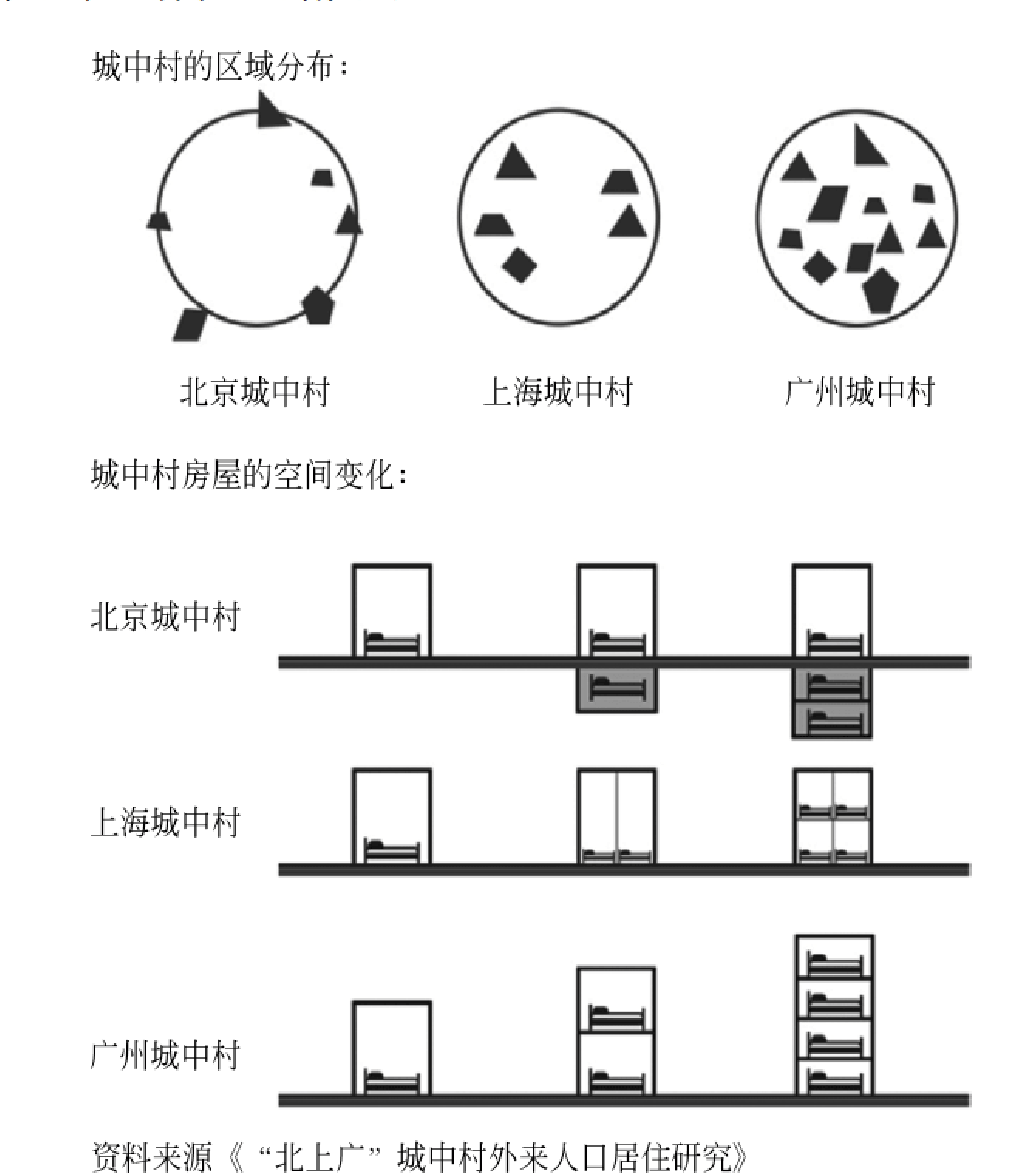


图 2.20　城中村区域分布及房屋空间变化

更为有趣的是，在大量外来人口涌入后，“北上广”三地城中村内房屋空间的变化。

北京多为不断下压的空间。在北京圈层的外扩中，内城的城中村逐步被拆迁。城郊村在形态上更多的呈现一种原始聚集村落形式，多为一层或两层的平房，每户拥有自己的院落房屋，部分有地下室。

上海则多是不断向内挤压空间。对于管治最为严格的上海，一方面迫于强硬的政策与监管，一方面又拥有异常旺盛的住房需求，所以只能在漫长的“等待拆迁”中通过内部挤压的方法“塞”进更多的人。村内原有的楼梯间、独立厨房、独立洗手间、院落等均被改造和分隔成住房。

相比较北京和上海，广州的城市监管较为松散，城中村多加向上加建房屋，表现出一种不断加建的空间。

4. 虽然可能并不幸福，但还是希望融入

青年们的人际交往状况又是如何？《中国流动人口发展报告》的结论是，北京、上海的外来青年中 6.3％、11.4％很少与人交往。

其中，上海的外来青年很少与取得上海户籍的同乡及本地人交往，将近 60％ 经常与同乡交往。而北京的外来青年更愿意与本地人来往，显示出更高的开放性和融入愿望，图 2.21 为外来青年人际交往状况。

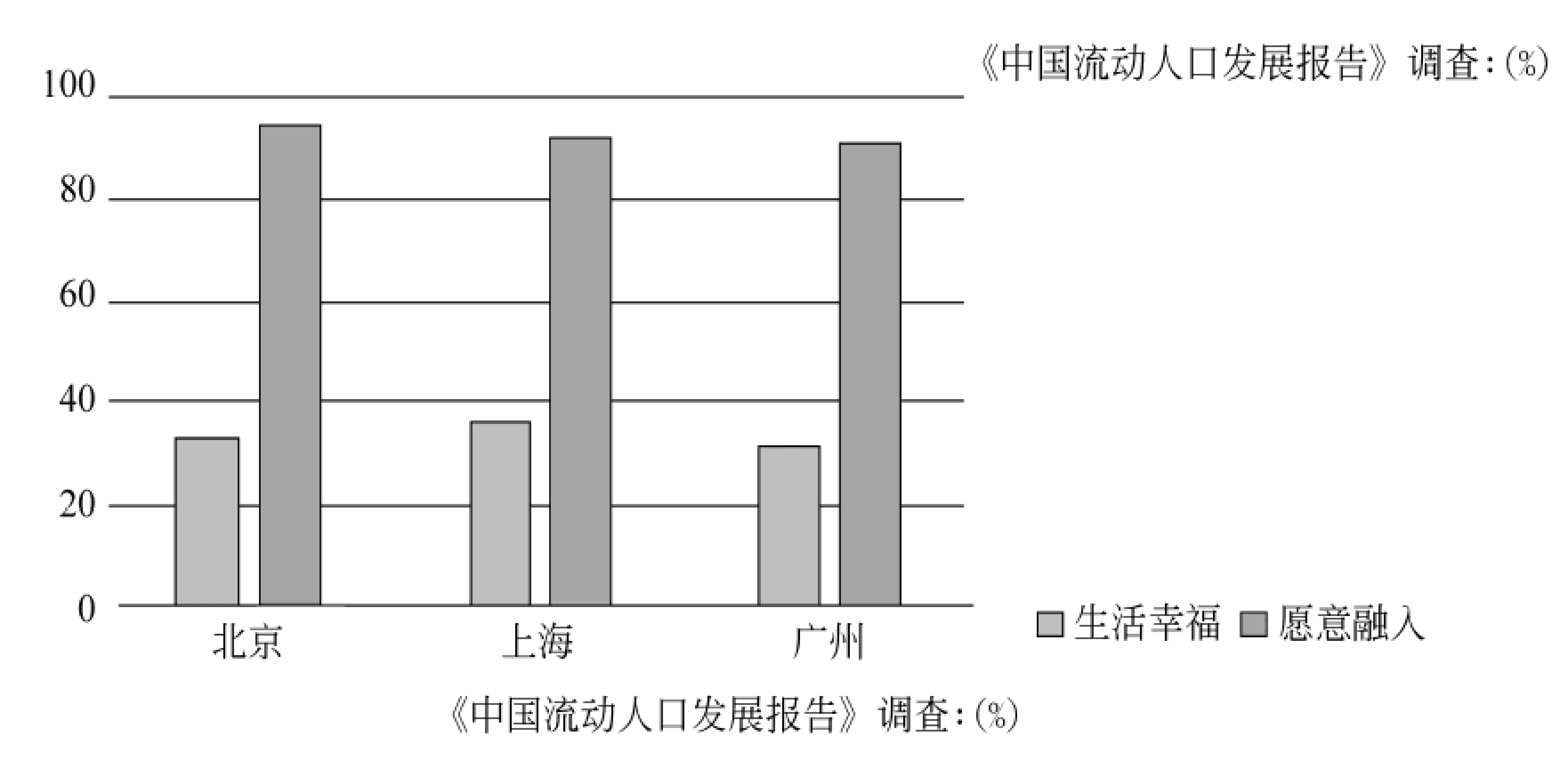


图 2.21 外来青年人际交往状况

如果问及在大都市生活“是否比在老家更幸福”，北京、上海的外来青年分别有 32.8％、35.8％的人回答肯定，略高于全国平均水平；而广州只有 28.4％的人感到幸福。但问及融入的意愿，“北上广”三地的外来青年均有超过 90％的人愿意融入。

资料来源：《中国流动人口发展报告》《“北上广”城中村外来人口居住研究》

习题与思考题

一、选择题

1. 目前，选用开源的虚拟化产品组建虚拟化平台，构建基于硬件的虚拟化层，可以选用(　　)。

A．Xen　B. VMware　C. Hyper-v　D. Citrix

2. 在云计算中，虚拟层主要包括(　　)。

A. 服务器虚拟化　B. 存储虚拟化

C. 网络虚拟化　D. 桌面虚拟化

3. Hadoop 项目包括(　　)。

A. Hadoop Distributed File System(HDFS)

B. Hadoop MapReduce 编程模型

C. Hadoop Streaming

D. Hadoop Common

4. 云计算的服务方式有(　　)。

A. IaaS　B. Raas　C. PaaS　D. SaaS

5. Amazon.com 公司通过(　　)计算云，可以让客户通过 Web Service 方式租用计算机来运行自己的应用程序。

A. S3　B. HDFS　C. EC2　D. GFS

6. 云是一个平台，是一个业务模式，给客户群体提供一些比较特殊的 IT 服务，分为(　　)三部分。(多选题)

A. 管理平台　B. 服务提供　C. 构建服务　D. 硬件更新

二、问答题

1. 什么是云计算？
2. 画图描述云计算系统的体系结构。
3. 简述云计算服务层次。
4. 云计算的核心技术有哪些？相互之间有什么关系？
5. 有哪几种典型的云计算系统？其分别应用在哪些方面？

第3章　大数据采集与预处理

3.1　大数据采集概念

足够的数据量是企业大数据战略建设的基础,因此数据采集就成了大数据分析的前站。采集是大数据价值挖掘重要的一环,其后的分析挖掘都建立在采集的基础上。大数据技术的意义确实不在于掌握规模庞大的数据信息,而在于对这些数据进行智能处理,从中分析和挖掘出有价值的信息,但前提是拥有大量的数据。

绝大多数的企业现在还很难判断,到底哪些数据未来将成为资产,通过什么方式将数据提炼为现实收入。对于这一点即便是大数据服务企业也很难给出确定的答案。但有一点是肯定的——大数据时代,谁掌握了足够的数据,谁就有可能掌握未来,现在的数据采集就是将来的资产积累。

数据的采集有基于物联网传感器的采集,也有基于网络信息的数据采集。比如在智能交通中,数据的采集有基于GPS的定位信息采集、基于交通摄像头的视频采集、基于交通卡口的图像采集、基于路口的线圈信号采集等。而在互联网上的数据采集是对各类网络媒介,如搜索引擎、新闻网站、论坛、微博、博客、电商网站等的各种页面信息和用户访问信息进行采集,采集的内容主要有文本信息、URL、访问日志、日期和图片等。之后我们需要把采集到的各类数据进行清洗、过滤、去重等各项预处理并分类归纳存储。

数据采集过程中涉及数据抽取、数据的清洗转换、数据的加载三个过程,其英文缩写为ETL(Extract、Transform、Load)。

数据采集的ETL工具负责将分布的、异构数据源中的不同种类和结构的数据如文本数据、关系数据以及图片、视频等非结构化数据等抽取到临时中间层后进行清洗、转换、分类、集成,最后加载到对应的数据存储系统如数据仓库或数据集市中,成为联机分析处理、数据挖掘的基础。

针对大数据的ETL工具同时又有别于传统的ETL处理过程,因为一方面大数据的体量巨大,另一方面数据的产生速度也非常快,比如一个城市的视频监控头、智能电表每一秒钟都在产生大量的数据,对数据的预处理需要实时快速,因此在ETL的架构和工具选择上,也会采用如分布式内存数据库、实时流处理系统等现代信息技术。

现代企业中存在各种不同的应用和各种数据格式及存储需求,但在企业之间、企业内部都存在条块分割、信息孤岛的现象,各个企业之间的数据不能实现可控的数据交换和共享,而且各个应用之间由于涉及开发技术和环境的限制也为企业的数据共享设置了障碍,阻碍了企业各个应用之间和数据交换和共享,也影响了企业对数据可控、数据管理、数据

安全方面的需求。为实现跨行业跨部门的数据整合，尤其是在智慧城市建设中，需要制定统一的数据标准、交换接口以及共享协议，这样不同行业、不同部门、不同格式的数据才能基于一个统一的基础进行访问、交换和共享。通过实现企业数据总线(EDS)，可以提供对企业应用中各类数据的存取功能，把企业数据的存取集成与企业的功能集成分离开来。

企业数据总线有效地创建了一层数据访问抽象层，使业务功能避开企业数据访问的细节。业务组件只需包含服务功能组件(用于实现现有服务功能)和数据访问组件(通过使用企业数据总线的方式)。通过企业数据总线这种方式，为企业的管理数据模型和应用系统数据模型间提供了一个统一的转换接口，并有效减少了各应用服务之间的耦合度。在大数据场景下，企业数据总线上会存在大量的同步的数据访问请求，总线上任何一个模块性能下降，都会大大影响总线功能，因此企业数据总线也需要采用大规模并发式、具备高可扩展性的实现方式。

3.2 数据采集来源

根据 MapReduce 产生数据的应用系统分类，大数据的采集主要有四种来源：管理信息系统、Web 信息系统、物理信息系统、科学实验系统。

1. 管理信息系统

管理信息系统是指企业、机关内部的信息系统，如事务处理系统、办公自动化系统，主要用于经营和管理，为特定用户的工作和业务提供支持。数据的产生既有终端用户的原始输入，也有系统的二次加工处理。系统的组织结构上是专用的，数据通常是结构化的。

2. Web 信息系统

Web 信息系统包括互联网上的各种信息系统，如社交网站、社会媒体、搜索引擎等，主要用于构造虚拟的信息空间，为广大用户提供信息服务和社交服务。系统的组织结构是开放式的，大部分数据是半结构化或无结构的。数据的产生者主要是在线用户。电子商务、电子政务是在 Web 上运行的管理信息系统。

3. 物理信息系统

物理信息系统是指关于各种物理对象和物理过程的信息系统，如实时监控、实时检测，主要用于生产调度、过程控制、现场指挥、环境保护等。系统的组织结构上是封闭的，数据由各种嵌入式传感设备产生的，可以是关于物理、化学、生物等性质和状态的基本测量值，也可以是关于行为和状态的音频、视频等多媒体数据。

4. 科学实验系统

科学实验系统实际上也属于物理信息系统，但其实验环境是预先设定的，主要用于研究和学术，数据是有选择的、可控的，有时可能是人工模拟生成的仿真数据。

在物理信息系统中，对于一个具体的物理对象，可采用不同观测手段，对其不同的属性(方面)进行测量，如测量一辆行驶汽车的尺寸、速度、路线、尾气、外观等，其观测结果为具有不同形式的数据，这些数据代表实体不同的模态，称为多模态(multi-modal)。对于

一个实体的多模态原始数据，需要做融合处理(data fusion)。在融合处理中，需要减少误差，保证数据的完整性和正确性。在高级的嵌入式系统或数据采集系统中，通常具有数据质量控制和数据融合处理功能。

从人-机-物三元世界观点看，管理信息系统和 Web 信息系统属于人与计算机的交互系统，物理信息系统属于物与计算机的交互系统。关于物理世界的原始数据，在人-机系统中，是通过人实现融合处理的；而在物-机系统中，需要通过计算机等装置做专门的处理。融合处理后的数据，被转换为规范的数据结构，输入并存储在专门的数据管理系统中，如文件或数据库，形成专门的数据集。

对于不同的数据集，可能存在不同的结构和模式，如文件、XML 树、关系表等，表现为数据的异构性(heterogeneity)。对多个异构的数据集，需要做进一步集成处理(data integration)或整合处理(data consolidation)，将来自不同数据集的数据收集、整理、清洗，转换后，生成到一个新的数据集，为后续查询和分析处理提供统一的数据视图。

3.3 大数据采集方法

3.3.1 大数据数据采集方面新方法

1. 系统日志采集方法

很多互联网企业都有自己的海量数据采集工具，多用于系统日志采集，如 Hadoop 的 Chukwa、Cloudera 的 Flume、Facebook 的 Scribe 等，这些工具均采用分布式架构，能满足每秒数百 MB 的日志数据采集和传输需求。

2. 网络数据采集方法：对非结构化数据的采集

网络数据采集是指通过网络爬虫或网站公开 API 等方式从网站上获取数据信息。该方法可以将非结构化数据从网页中抽取出来，将其存储为统一的本地数据文件，并以结构化的方式存储。它支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。

除了网络中包含的内容之外，对于网络流量的采集可以使用 DPI 或 DFI 等带宽管理技术进行处理。

3. 其他数据采集方法

对于企业生产经营数据或学科研究数据等保密性要求较高的数据，可以通过与企业或研究机构合作，使用特定系统接口等相关方式采集数据。

3.3.2 网页数据采集方法

互联网网页数据具有分布广、格式多样、非结构化等大数据的典型特点，我们需要有针对性地对互联网网页数据进行采集、转换、加工和存储，尤其在网页数据的采集和处理方面，存在亟须突破的若干关键技术。

传统的数据挖掘、分析处理方法和工具，在非结构化、高速化的大数据处理要求面前

显得过于乏力，需要创新开发适应新型大数据处理需求的数据挖掘和数据处理方法。

互联网网页数据是大数据领域的一个重要组成部分，是互联网公司和金融机构获取用户消费、交易、产品评价信息以及其他社交信息等数据的重要途径，为互联网和金融服务创新提供了丰富的数据基础，因此，对互联网网页的大数据处理流程和技术进行探索具有重要意义。

1. 网页大数据采集的基本流程

互联网网页数据采集就是获取互联网中相关网页内容的过程，并从中抽取出用户所需要的属性内容。互联网网页数据处理，就是对抽取出来的网页数据进行内容和格式上的处理，进行转换和加工，使之能够适应用户的需求，并将之存储下来，以供后用。

网络爬虫是一个自动提取网页的程序，它为搜索引擎从万维网上下载网页，是搜索引擎的重要组成部分。传统爬虫从一个或若干初始网页的 URL 开始，获得初始网页上的 URL，在抓取网页的过程中，不断从当前页面上抽取新的 URL 放入队列，直到满足系统的一定停止条件。

聚焦爬虫的工作流程较为复杂，需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的 URL 队列。然后，它将根据一定的搜索策略从队列中选择下一步要抓取的网页 URL，并重复上述过程，直到达到系统的某一条件时停止。

另外，所有被爬虫抓取的网页将会被系统存储，进行一定的分析、过滤，并建立索引，以便之后的查询和检索；对于聚焦爬虫来说，这一过程所得到的分析结果还可能对以后的抓取过程给出反馈和指导。网络爬虫自动提取网页的过程见图 3.1。

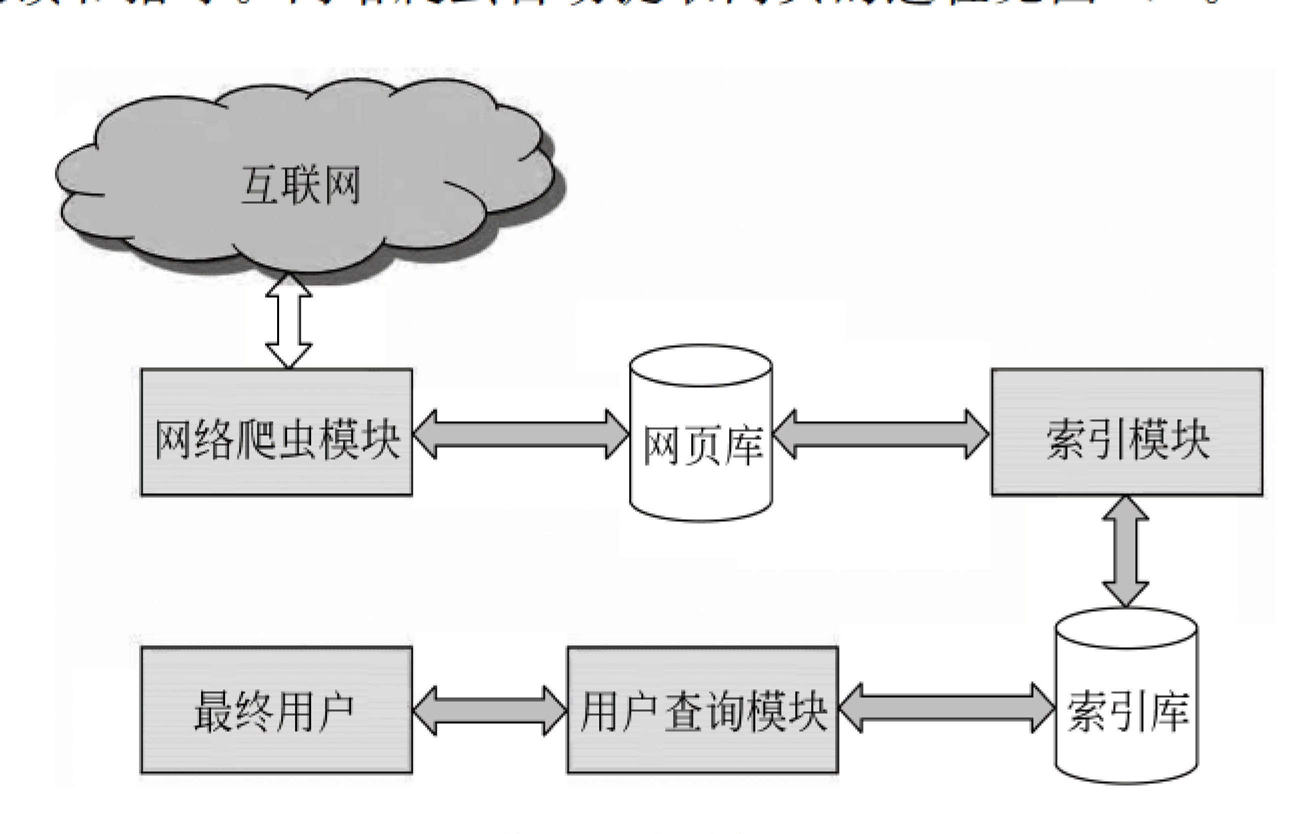


图 3.1　网络爬虫自动提取网页的过程

2. 网页数据采集工作过程

1）工作过程描述

采集的目的就是把对方网站上网页中的某块文字或者图片等资源下载到自己的站网上，这个过程需要做如下配置工作：下载网页配置，解析网页配置，修正结果配置，数据输出配置。如果数据符合自己要求，修正结果这步可省略。配置完毕后，把配置形成任务（任务以 XML 格式描述），采集系统按照任务的描述开始工作，最终把采集到的结果存储到网站服务器上。

2）工作流程

整个数据采集过程的基本步骤如下：

(1) 将需要抓取数据的网站的 URL 信息(Site URL)写入 URL Queue；

(2) 爬虫从 URL 队列中获取需要抓取数据的网站的 Site URL 信息；

(3) 获取某个具体网站的网页内容；

(4) 从网页内容中抽取出该网站正文页内容的链接地址；

(5) 从数据库中读取已经抓取过内容的网页地址(Spider URL)；

(6) 过滤 URL。将当前的 URL 和已经抓取过的 URL 进行比较；

(7) 如果该网页地址没有被抓取过，则将该地址写入(Spider URL)数据库；如果该地址已经被抓取过，则放置对这个地址的抓取操作；

(8) 获取该地址的网页内容，并抽取出所需属性的内容值；

(9) 将抽取的网页内容写入数据库。

数据采集工作流程图如图 3.2 所示。

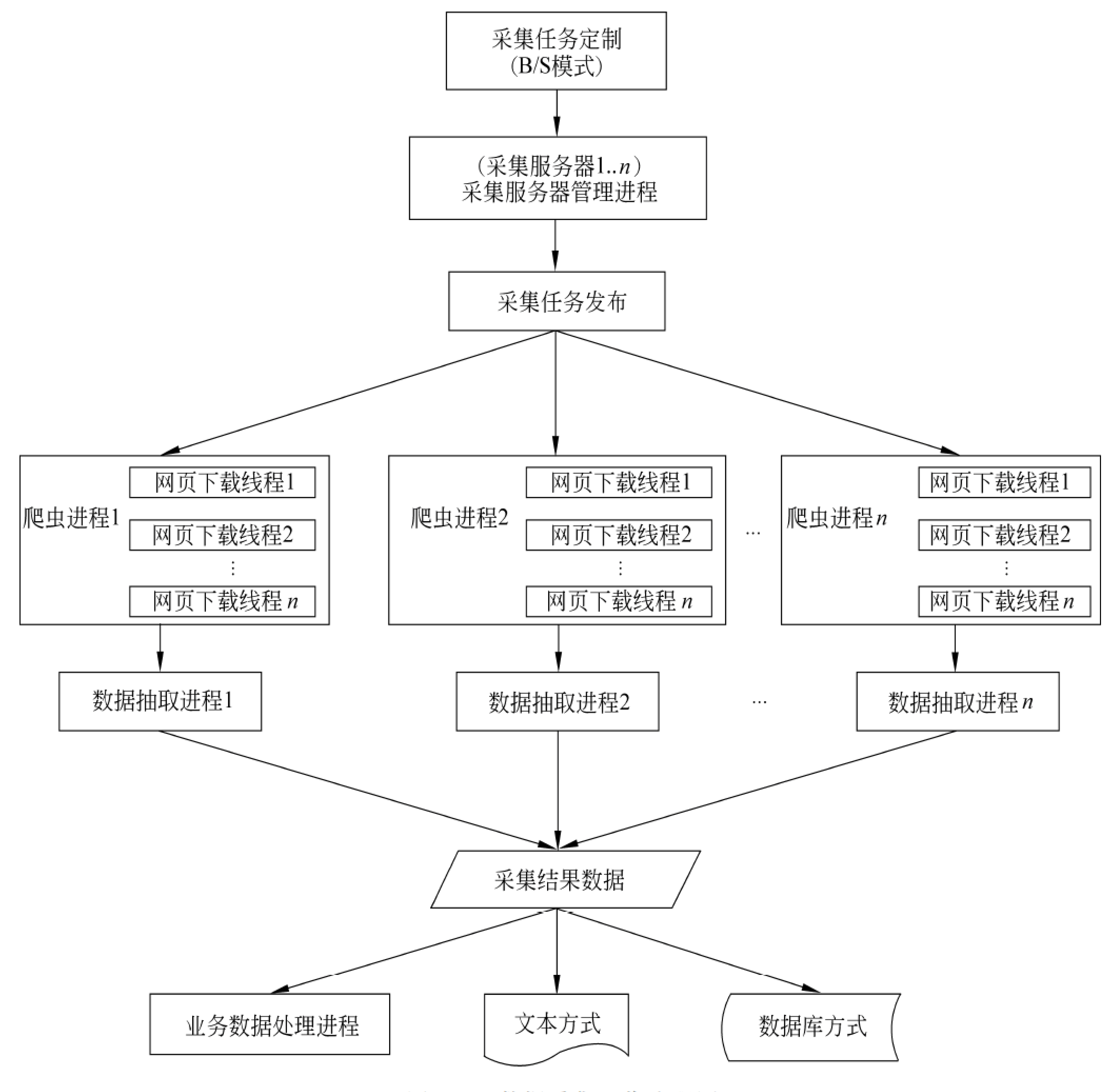


图 3.2 数据采集工作流程图

相应的网页内容提取、数据采集与数据处理逻辑如图 3.3 所示。

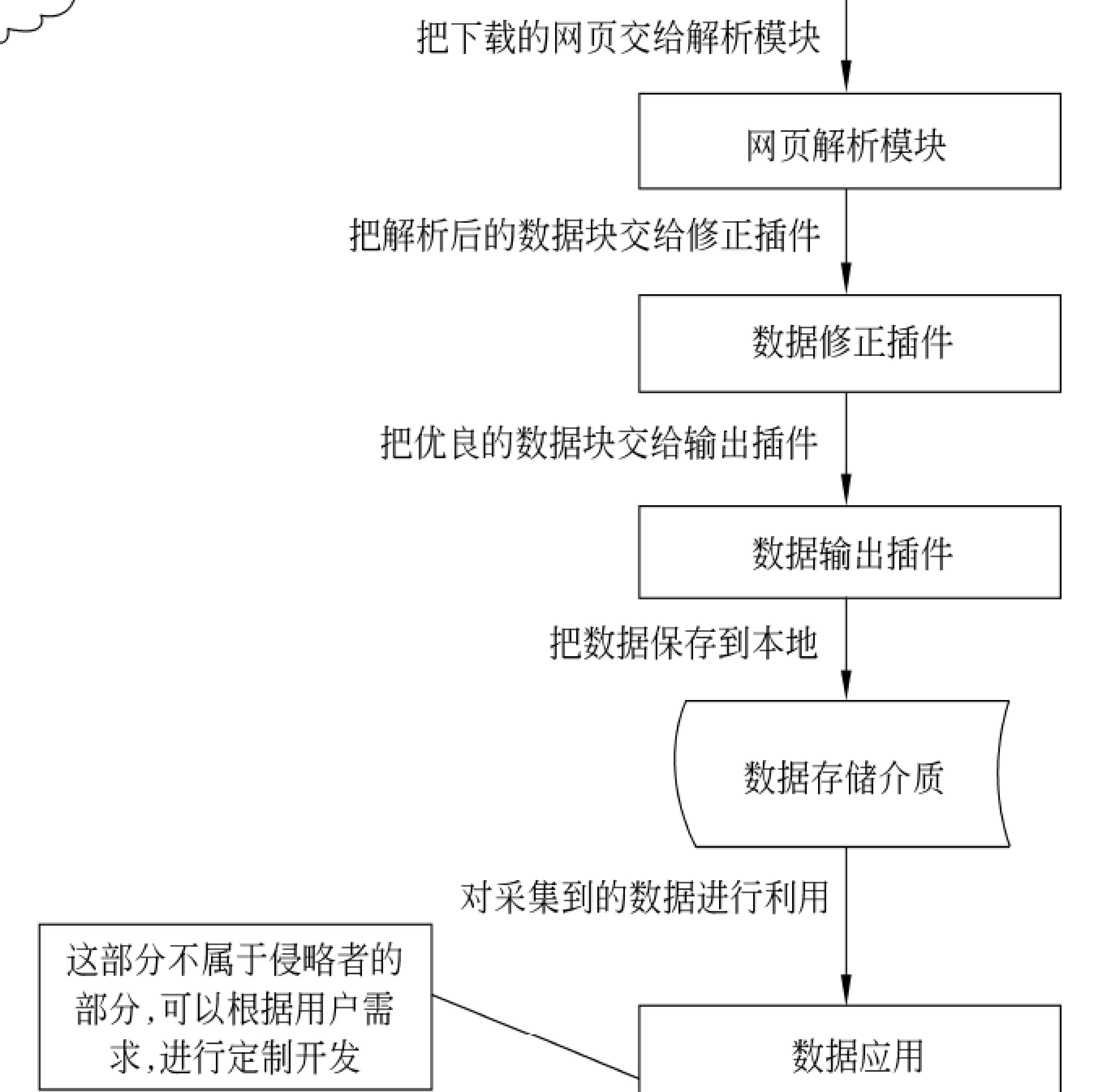


图 3.3　网页内容提取、数据采集与数据处理逻辑

3.3.3　Web 信息数据自动采集

Web 可以说是目前最大的信息系统,其数据具有海量、多样、异构、动态变化等特性。因此人们要准确迅速地获得自己所需要的数据越来越难,尽管目前有各种搜索引擎,但是搜索引擎在数据的查全率考虑较多,而查准率不足,而且很难进一步挖掘深度数据。因此人们开始研究如何更进一步获取互联网上某一个特定范围的数据,从信息搜索到知识发现。

1. Web 数据自动采集相关概念

Web 数据自动采集涉及 Web 数据挖掘(Web Data Mining)、Web 信息检索(Web Information Revival)、信息提取(Information Extraction)、搜索引擎(Search Engine)等概念和技术。Web 数据挖掘与这些概念密切相关,但又有所区别。

1) Web 数据自动采集与挖掘

所谓 Web 数据自动采集与挖掘,是指从大量非结构化、异构的 Web 信息资源中发现

有效的、新颖的、潜在可用的及最终可以理解的知识(包括概念、模式、规则、规律、约束及可视化等形式)的非平凡过程。

2) Web 数据自动采集与搜索引擎

Web 数据自动采集与搜索引擎有许多相似之处,比如它们都利用了信息检索的技术。但是两者侧重点不同,搜索引擎主要由网络爬虫(Web Scraper)、索引数据库和查询服务三个部分组成。爬虫在网上的漫游是无目的性的,只是尽量发现比较多的内容。查询服务尽可能多地返回结果,但不关心结果是否符合用户的习惯专业背景等。而 Web 数据自动采集主要针对某个具体行业,提供面向领域,个性化的信息挖掘服务。

3) Web 数据自动采集与信息提取

信息提取(Information Extraction)是近年来新兴的一个概念。信息提取是面向不断增长和变化的,某个具体领域的文献特定的查询,这种查询是长期的或者持续的。与传统搜索引擎是基于关键字查询的不同,信息提取基于查询。不仅要包含关键字,还要匹配各个实体之间的关系。信息提取是技术上的概念。Web 数据自动采集很大程度要依赖于信息提取的技术,实现长期的、动态的追踪。

4) Web 数据自动采集与 Web 信息检索

信息检索即从大量的 Web 文献集合 C 中,找到与给定查询 q 相关的,数目相当的文献子集 S。如果将 q 看作输入,S 看作输出,那么 Web 信息检索的过程就是一个输入到输出的映像:

$$\xi:(C:q)\to S3$$

而 Web 数据自动采集不是直接将 Web 文献集合的子集直接输出给用户,还要进一步的分析处理,查重去噪,整合数据等。尽量将半结构化甚至非结构化的数据变为结构化的数据,然后以统一的格式呈现给用户。

因此,Web 数据自动采集是 Web 数据挖掘的一个重要组成部分,它利用了 Web 数据检索、信息提取的技术,弥补了搜索引擎缺乏针对性和专业性,不能实现数据的动态跟踪与监测的缺点,是一个非常有前景的领域。

2. 数据采集的关键技术——链接过滤

链接过滤的实质就是判断一个链接(当前链接)是不是在一个链接集合(已经抓取过的链接)里面。在对网页大数据的采集中,可以采用布隆过滤器来实现对链接的过滤。

布隆过滤器(Bloom Filter)的基本思想是:当一个元素被加入集合时,通过 K 个散列函数将这个元素映射成一个位数组中的 K 个点,把它们置为 1。检索时,我们只要看看这些点是不是都是 1(大约)就知道集合中有没有它了:如果这些点有任何一个 0,则被检元素一定不在;如果都是 1,则被检元素很可能在。

布隆过滤器在空间和时间方面都有巨大的优势:

(1) 在复杂度方面,布隆过滤器存储空间和插入/查询时间都是常数(即复杂度为 $O(k)$);

(2) 在关系方面,散列函数相互之间没有关联关系,方便由硬件并行实现;

(3) 在存储方面,布隆过滤器不需要存储元素本身,在某些对保密要求非常严格的场合有优势。

布隆过滤器的具体实现方法是,已经抓取过的每个 URL,经过 k 个 hash 函数的计算,得出 k 个值,再和一个巨大 bit 数组的这 k 个位置的元素对应起来(这些位置数组元素的值被设置为 1)。在需要判断某个 URL 是否被抓取过时,先用 k 个 hash 函数对该 URL 计算出 k 个值,然后查询巨大的 bit 数组内这 k 个位置上的值,如果全为 1,则是已经被抓取过,否则没有被抓取过。

3. Web 引擎和通用搜索引擎的差别

Web 结构化信息抽取就是将网页中的非结构化数据按照一定的需求抽取成结构化数据。是垂直搜索。

Web 引擎和通用搜索引擎比较大的差别,例如:

(1) 比较购物搜索需要在抓取网页后,对网页中的商品信息进行抽取,抽取出商品名称、价格、简介……甚至可以进一步将笔记本简介细分成"品牌、型号、CPU、内存、硬盘、显示屏、……"

(2) 房产信息搜索应该抽取出:类型、地域、地址、房型、面积、装修情况、租金、联系人、联系电话,公司企业信息搜索应该抽取出:公司名称、地址、电话、联系人。

结构化信息抽取有两种方式可以实现,比较简单的是模板方式,还有一种是对网页不依赖的网页库级的结构化信息抽取方式。

(1) 模板方式。

模板方式是事先对特定的网页进行配置模板,抽取模板中设置好的需要的信息,可以针对有限个网站的信息进行精确的采集。

特点:简单、精确、技术难度低、方便快速部署。

缺点:需要针对每一个信息源的网站模板进行单独的设定,在信息源多样性的情况下维护量巨大是指不可完成的维护量。所以这种方式适合少量信息源的信息处理,不是搜索引擎级的应用,很难满足用户对查全率的需求。

(2) 网页库级的结构化信息抽取方式。

网页库结构化信息抽取是采用页面结构分析与智能结点分析转换的方法,自动抽取结构化的数据。

特点:可对任意的正常网页进行抽取,完全自动化,不用对具体网站事先生成模板,对每个网页自动实时得生成抽取规则,完全不需要人工干预。智能抽取准确率高,不是机械的匹配,采用智能分析技术,准确率能达到 98%以上。能保证较快处理速度,由于采用页面的智能分析技术,先去除了垃圾块,降低分析的压力,使处理速度大大提高。通用性较好,易于维护,只需设定参数、配置相应的特征就能改进相应的抽取性能;一般的非专业人员经过简单培训就能维护。

缺点:技术难度高,前期研发成本高,周期长。适合网页库级别结构化数据采集和搜索的高端应用。

3.4 导入/预处理

3.4.1 大数据导入/预处理的过程

大数据处理是将业务系统的数据经过抽取、清洗转换之后加载到数据仓库的过程,目的是将企业中的分散、零乱、标准不统一的数据整合到一起,为企业的决策提供分析的依据。数据抽取、清洗与转换是大数据处理最重要的一个环节,通常情况下会花掉整个项目的 1/3 的时间。

数据的抽取是从各个不同的数据源抽取到处理系统中,在抽取的过程中需要挑选不同的抽取方法,尽可能提高运行效率。花费时间最长的是清洗、转换的部分,一般情况下这部分工作量是整个过程的 2/3。数据的加载一般在数据清洗完后直接写入数据仓库中去。

数据抽取、清洗与转换的实现有多种方法,常用的有三种:第一种是借助工具如 Oracle 的 OWB、SQL Server 2000 的 DTS、SQL Server 2005 的 SSIS 服务等实现;第二种是 SQL 方式实现;第三种是工具和 SQL 相结合。前两种方法各有优缺点,借助工具可以快速地建立起工程,屏蔽复杂的编码任务,提高速度,降低难度,但是欠缺灵活性。SQL 的方法优点是灵活,提高运行效率,但是编码复杂,对技术要求比较高。第三种综合了前面两种的优点,极大地提高了开发速度和效率。

1. 数据的抽取

数据的抽取需要在调研阶段做大量工作,首先要搞清楚以下几个问题:数据是从几个业务系统中来?各个业务系统的数据库服务器运行什么数据库管理系统(DBMS)?是否存在手工数据?手工数据量有多大?是否存在非结构化的数据?等等类似问题,当收集完这些信息之后才可以进行数据抽取的设计。

1)与存放数据仓库(Data Warehouse,DW)的数据库系统相同的数据源处理方法

这一类数源在设计比较容易,一般情况下,DBMS(包括 SQL Server、Oracle)都会提供数据库链接功能,在 DW 数据库服务器和原业务系统之间建立直接的链接关系就可以写 Select 语句直接访问。

2)与 DW 数据库系统不同的数据源的处理方法

这一类数据源一般情况下也可以通过 ODBC 的方式建立数据库链接,如 SQL Server 和 Oracle 之间。如果不能建立数据库链接,可以通过两种方式完成:一种是通过工具将源数据导出成.txt 或者是.xls 文件,然后再将这些源系统文件导入到 ODS 中;另外一种方法通过程序接口来完成。

3)对于文件类型数据源(.txt、.xls),可以培训业务人员利用数据库工具将这些数据导入到指定的数据库,然后从指定的数据库抽取。或者可以借助工具实现,如 SQL Server 2005 的 SSIS 服务的平面数据源和平面目标等组件导入 ODS 中去。

4)增量更新问题

对于数据量大的系统,必须考虑增量抽取。一般情况,业务系统会记录业务发生的时

间，可以用作增量的标志，每次抽取之前首先判断 ODS 中记录最大的时间，然后根据这个时间去业务系统取出大于这个时间的所有记录。利用业务系统的时间戳，一般情况下，业务系统没有或者部分有时间戳。

2. 数据的清洗转换

一般情况下，数据仓库分为 ODS、DW 两部分，通常的做法是从业务系统到 ODS 做清洗，将脏数据和不完整数据过滤掉，再从 ODS 到 DW 的过程中转换，进行一些业务规则的计算和聚合。

1）数据清洗

数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉还是由业务单位修正之后再进行抽取。不符合要求的数据主要是有不完整的数据、错误的数据和重复的数据三大类。

（1）不完整的数据。

其特征是一些应该有的信息缺失，如供应商的名称、分公司的名称、客户的区域信息缺失、业务系统中主表与明细表不能匹配等。需要将这一类数据过滤出来，按缺失的内容分别写入不同 Excel 文件向客户提交，要求在规定的时间内补全，补全后才写入数据仓库。

（2）错误的数据。

其产生原因是业务系统不够健全，在接收输入后没有进行判断就直接写入后台数据库造成的，比如数值数据输成全角数字字符、字符串数据后面有一个回车、日期格式不正确、日期越界等。这一类数据也要分类，对于类似于全角字符、数据前后有不面见字符的问题只能以 SQL 的方式找出来，然后要求客户在业务系统修正之后抽取；日期格式不正确的或者是日期越界的这一类错误会导致 ETL 运行失败，这一类错误需要去业务系统数据库用 SQL 的方式挑出来，交给业务主管部门要求限期修正，修正之后再抽取。

（3）重复的数据。

在维度表中比较常见，将重复的数据的记录所有字段导出来，让客户确认并整理。

数据清洗是一个反复的过程，不可能在几天内完成，只有不断地发现问题，解决问题。对于是否过滤、是否修正一般要求客户确认；对于过滤掉的数据，写入 Excel 文件或者将过滤数据写入数据表，在 ETL 开发的初期可以每天向业务单位发送过滤数据的邮件，促使他们尽快修正错误，同时也可以作为将来验证数据的依据。数据清洗需要注意的是不要将有用的数据过滤掉了，对于每个过滤规则认真进行验证，并要用户确认才行。

2）数据转换

数据转换的任务主要是进行不一致的数据转换、数据粒度的转换和一些商务规则的计算。

（1）不一致数据转换

这个过程是一个整合的过程，将不同业务系统的相同类型的数据统一，比如同一个供应商在结算系统的编码是 XX0001，而在 CRM 中编码是 YY0001，这样在抽取过来之后

统一转换成一个编码。

(2) 数据粒度的转换

业务系统一般存储非常明细的数据,而数据仓库中的数据是用来分析的,不需要非常明细的数据,一般情况下,会将业务系统数据按照数据仓库粒度进行聚合。

(3) 商务规则的计算

不同的企业有不同的业务规则和不同的数据指标,这些指标有的时候不是简单的加加减减就能完成,这个时候需要在ETL中将这些数据指标计算好了之后存储在数据仓库中,供分析使用。

3.4.2 数据清洗

科研工作者、工程师、业务分析者都要和数据打交道,数据分析在他们的工作中是一项核心任务。这么不仅仅针对"大数据"的从业者,即使是你笔记本硬盘上的数据也值得分析。数据分析的第一步是洗数据,原始数据可能有各种不同的来源,包括:

(1) Web服务器的日志。

(2) 某种科学仪器的输出结果。

(3) 在线调查问卷的导出结果。

(4) 政府数据。

(5) 企业顾问准备的报告。

在理想世界中,所有记录都应该是整整齐齐的格式,并且遵循某种简洁的内在结构。但是实际中可不是这样。所有这些数据的共同点是:你绝对料想不到它们的各种怪异的格式。数据给你了,那就要处理,但这些数据可能经常是:

(1) 不完整的(某些记录的某些字段缺失)。

(2) 前后不一致(字段名和结构前后不一)。

(3) 数据损坏(有些记录可能会因为种种原因被破坏)。

因此,你必须经常维护你的清洗程序来清洗这些原始数据,把它们转化成易于分析的格式,通常称为数据清洗(data wrangling)。接下来会介绍一些关于如何有效清洗数据,所有介绍的内容都可以由任意编程语言实现。

1. 不符合要求的数据

数据清洗从名字上也看得出就是把"脏"的"洗掉"。因为数据仓库中的数据是面向某一主题的数据的集合,这些数据从多个业务系统中抽取而来而且包含历史数据,这样就避免不了有的数据是错误数据、有的数据相互之间有冲突,这些错误的或有冲突的数据显然是我们不想要的,称为"脏数据"。我们要按照一定的规则把"脏数据""洗掉",这就是数据清洗。而数据清洗的任务是过滤那些不符合要求的数据,将过滤的结果交给业务主管部门,确认是否过滤掉还是由业务单位修正之后再进行抽取。

不符合要求的数据主要是有不完整的数据、错误的数据、重复的数据三大类,见3.4.1节所述。

2. 数据清洗

洗数据的程序肯定会经常崩溃。这很好,因为每一次崩溃都意味着你这些糟糕的数

据又跟你最初的假设相悖了。反复地改进你的断言直到能成功走通。但一定要尽可能让其保持严格，不要太宽松，要不然可能达不到你要的效果。最坏的情况不是程序走不通，而是走出来不是你要的结果。

以下是一些数据清洗的经验。

1）不要默默地跳过记录

原始数据中有些记录是不完整或者损坏的，所以洗数据的程序只能跳过。默默地跳过这些记录不是最好的办法，因为你不知道什么数据遗漏了。因此，这样做更好：

（1）打印出 warning 提示信息，这样你就能够过后再去寻找什么地方出错了。

（2）记录总共跳过了多少记录，成功清洗了多少记录。这样做能够让你对原始数据的质量有个大致的感觉，比如，如果只跳过了 0.5%，这还说得过去；但是如果跳过了 35%，那就该看看这些数据或者代码存在什么问题了。

2）使用 Set 或者 Counter 把变量的类别以及类别出现的频次存储起来

数据中经常有些字段是枚举类型的。例如，血型只能是 A、B、AB 或者 O。用断言来限定血型只能是这 4 种之一虽然挺好，但是如果某个类别包含多种可能的值，尤其是当有的值你可能始料未及的话，就不能用断言了。这时候，采用 counter 这种数据结构来存储就会比较好用。这样做你就可以：

（1）对于某个类别，假如碰到了始料未及的新取值时，就能够打印一条消息提醒你一下。

（2）洗完数据之后供你反过头来检查。例如，假如有人把血型误填成 C，那回过头来就能轻松发现了。

3）断点清洗

如果你有大量的原始数据需要清洗，要一次清洗完可能需要很久，有可能是 5 分钟、10 分钟、一小时，甚至是几天。实际当中，经常在洗到一半的时候突然崩溃了。

假设你有 100 万条记录，你的清洗程序在第 325 392 条因为某些异常崩溃了，你修改了这个 bug，然后重新清洗，这样的话，程序就得重新从 1 清洗到 325 391，这是在做无用功。其实可以这么做：

第一步，让你的清洗程序打印出来当前在清洗第几条，这样，如果崩溃了，你就能知道处理到哪条时崩溃了。

第二步，让你的程序支持在断点处开始清洗，这样当重新清洗时，你就能从 325 392 条直接开始。重洗的代码有可能会再次崩溃，你只要再次修正 bug，然后从再次崩溃的记录开始就行了。

当所有记录都清洗结束之后，再重新清洗一遍，因为后来修改 bug 后的代码可能会对之前的记录的清洗带来一些变化，两次清洗保证万无一失。但总的来说，设置断点能够节省很多时间，尤其是当你在 debug 的时候。

4）在一部分数据上进行测试

不要尝试一次性清洗所有数据。当你刚开始写清洗代码和 debug 的时候，在一个规模较小的子集上进行测试，然后扩大测试的这个子集再测试。这样做的目的是能够让清洗程序很快完成测试集上的清洗，例如几秒，这样会节省你反复测试的时间。

但是要注意,这样做的话,用于测试的子集往往不能涵盖到一些特别记录。

5)把清洗日志打印到文件中

当运行清洗程序时,把清洗日志和错误提示都打印到文件当中,这样就能轻松使用文本编辑器来查看它们了。

6)可选:把原始数据一并存储下来

当你不用担心存储空间的时候,这一条经验还是很有用的。这样做能够让原始数据作为一个字段保存在清洗后的数据当中,在清洗完之后,如果你发现哪条记录不对了,就能够直接看到原始数据长什么样子,方便你解决问题(debug)。

不过,这样做的坏处就是需要消耗双倍的存储空间,并且让某些清洗操作变得更慢。所以这一条只适用于效率允许的情况下。

7)最后一点,验证清洗后的数据

记得写一个验证程序来验证清洗后得到的干净数据是否跟预期的格式一致。你不能控制原始数据的格式,但是能够控制干净数据的格式。所以,一定要确保干净数据的格式是符合你预期的格式的。

这一点其实是非常重要的,因为完成了数据清洗之后,接下来就会直接在这些干净数据上进行下一步工作了。因此,在你开始数据分析之前要确保数据是足够干净的。否则,你可能会得到错误的分析结果,到那时候,就很难再发现很久之前的数据清洗过程中犯的错了。

3.4.3 数据采集(ETL)技术

随着信息化进程的推进,人们对数据资源整合的需求越来越明显。但面对分散在不同地区、种类繁多的异构数据库进行数据整合并非易事,要解决冗余、歧义等脏数据的清洗问题,仅靠手工进行不但费时费力,质量也难以保证;另外,数据的定期更新也存在困难。如何实现业务系统数据整合,是摆在大数据面前的难题。ETL数据转换系统为数据整合提供了可靠的解决方案。

ETL是Extraction-Transformation-Loading的缩写,中文名称为数据提取、转换和加载。ETL负责将分布的、异构数据源中的数据如关系数据、平面数据文件等抽取到临时中间层后进行清洗、转换、集成,最后加载到数据仓库或数据集市中,成为联机分析处理、数据挖掘的基础。它可以批量完成数据抽取、清洗、转换、装载等任务,不但满足了人们对种类繁多的异构数据库进行整合的需求,同时可以通过增量方式进行数据的后期更新。

ETL体系结构体现了主流ETL产品的主要组成部分,其体系结构如图3.4所示。

ETL过程中的主要环节就是数据抽取、数据转换和加工、数据装载。为了实现这些功能,各个ETL工具一般会进行一些功能上的扩充,例如工作流、调度引擎、规则引擎、脚本支持、统计信息等。

1. 数据抽取

数据抽取是从数据源中抽取数据的过程。实际应用中,不管数据源采用的是传统关

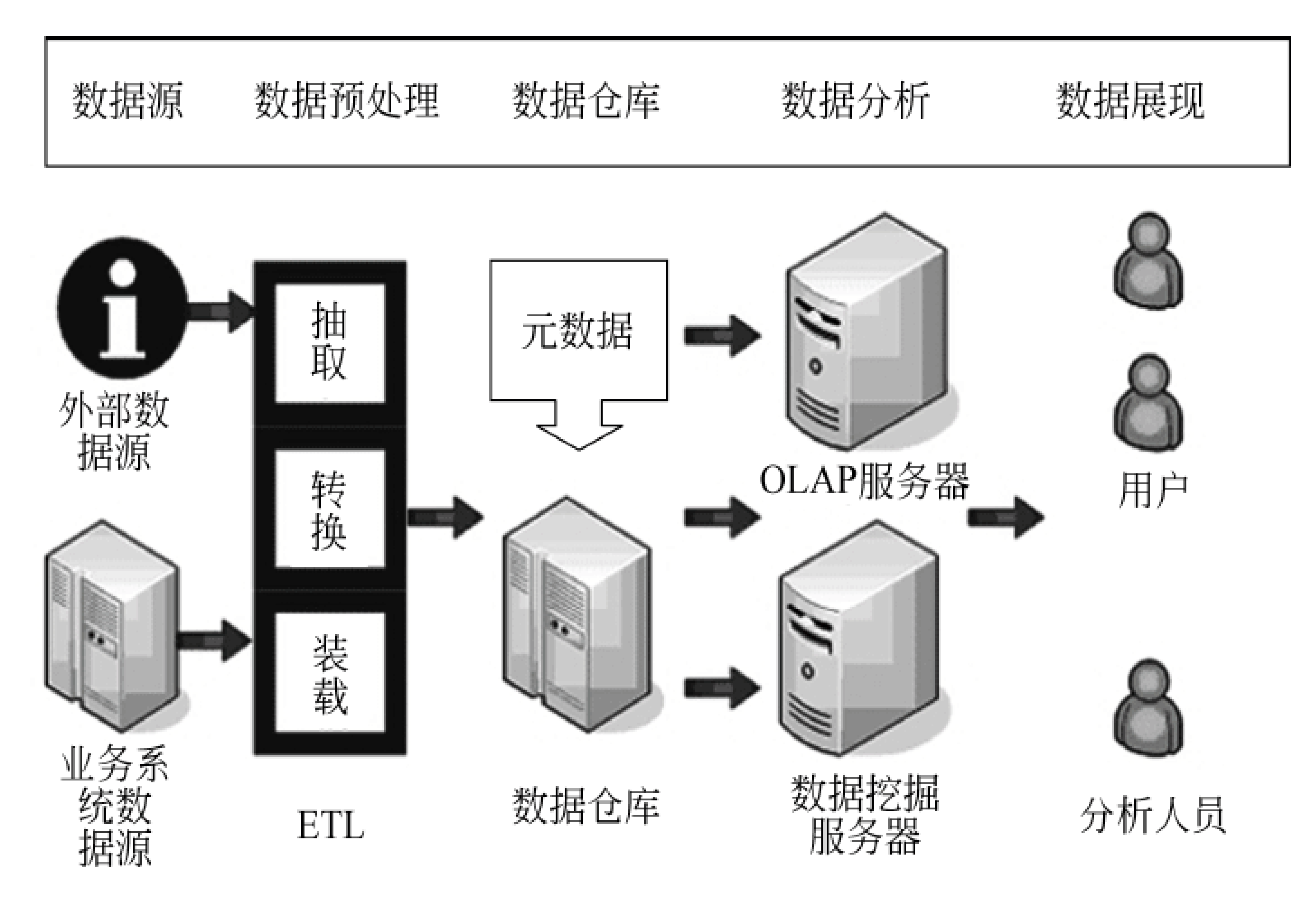


图 3.4　ETL 体系结构

系数据库还是新兴的 NoSQL 数据库，数据抽取一般有以下几种方式。

1）全量抽取

全量抽取指的是 ETL 在集成端进行数据的初始化时，首先由业务人员或相关的操作人员定义抽取策略，选定抽取字段和定义规则后，由设计人员进行程序设计；将数据进行处理后，直接读取整个工作表中的数据作为抽取的内容，类似于数据迁移，是 ETL 过程中最简单的步骤，其简单性使其主要适用于处理一些对用户非常重要的数据表。

2）增量抽取

增量抽取主要发生在全量抽取之后。全量抽取之后，对上次抽取过的数据源表中新增的或被修改的数据进行抽取，称为增量抽取。增量抽取可以减少对抽取过程中的数据量，提高抽取速度和效率，减少网络流量，同时，增量抽取的实现，对异构数据源和数据库中数据的变化有个准确的把握。信息抽取不是仅仅从大量的文献集或数据集中找出适合用户需要的那篇文献或部分内容，而是抽取出真正适合用户需要的相关信息片段，提供给用户，并找出这些信息与原文献直接的参考对照。

2. 数据转换和加工

从数据源中抽取的数据不一定完全满足目的库的要求，例如数据格式的不一致、数据输入错误、数据不完整等等，还要对抽取出的数据进行数据转换和加工。

数据转换是真正将源数据库中的数据转换为目标数据的关键步骤，在这个过程中，通过对数据的合并、汇总、过滤以及重新格式化和再计算等，从而将操作型数据库中的异构数据转换成用户所需要的形式。数据的转换和加工可以在 ETL 引擎中进行，也可以在数据抽取过程中利用数据库的特性同时进行。

1）ETL 引擎中的数据转换和加工

ETL 引擎中一般以组件化的方式实现数据转换。常用的数据转换组件有字段映射、数据过滤、数据清洗、数据替换、数据计算、数据验证、数据加解密、数据合并、数据拆分等。这些组件如同一条流水线上的一道道工序，它们是可插拔的，且可以任意组装，各组件之间通过数据总线共享数据。有些 ETL 工具还提供了脚本支持，使得用户可以以一种编程

的方式定制数据的转换和加工行为。

2）在数据库中进行数据加工

关系数据库本身已经提供了强大的 SQL、函数来支持数据的加工，如在 SQL 查询语句中添加 where 条件进行过滤、查询中重命名字段名与目的表进行映射、使用 case 条件判断等等。相比在 ETL 引擎中进行数据转换和加工，直接在 SQL 语句中进行转换和加工更加简单清晰，性能更高。对于 SQL 语句无法处理的可以交由 ETL 引擎处理。

3. 数据装载

将转换和加工后的数据装载到目的库中通常是 ETL 过程的最后步骤。装载数据的最佳方法取决于所执行操作的类型以及需要装入多少数据。当目的库是关系数据库时，一般来说有两种装载方式。

1）SQL 装载

直接 SQL 语句进行 insert、update、delete 操作。

2）采用批量装载方法

如 bcp、bulk、关系数据库特有的批量装载工具或 API。

大多数情况下会使用第一种方法，因为它们进行了日志记录并且是可恢复的。但是，批量装载操作易于使用，并且在装入大量数据时效率较高。使用哪种数据装载方法取决于业务系统的需要。

3.4.4 基于大数据的数据预处理

毫无疑问，数据预处理在整个数据挖掘流程中有非常重要的地位，可以说 60％甚至更多的时间和资源都花费在数据预处理上了。

传统背景下数据预处理更多的是对数据库的清洗，可能是 MySQL、Oracle 之类的数据，这些数据有着比较固定的模式，数据维度也不是很多，而且每一维度的数据类型（离散、连续数值、类标）以及包含的信息都能很明确。

而大数据背景下的数据预处理更倾向于对数据仓库的清洗，首先数据都是异源（各种数据来源），这个要统一起来就有大的工作量；其次数据可能没有固定的结构，或者称为非结构化数据，比如文本；第三，就是所谓的数据量大，大到单机程序或者小的分布式集群无法在给定时间范围内处理完毕；第四，就是数据量太大导致很多有用的信息被噪声淹没，甚至都不知道这些数据能干什么，分不清主次！

1. 为什么要预处理数据

（1）现实世界的数据是“肮脏”的（不完整、含噪声、不一致）。

（2）没有高质量的数据，就没有高质量的挖掘结果（高质量的决策必须依赖于高质量的数据；数据仓库需要对高质量的数据进行一致地集成）。

（3）原始数据中存在的问题包括存在不一致（数据内含出现不一致情况）、重复、不完整（没有感兴趣的属性）、含噪声（数据中存在着错误）、高维度或异常（偏离期望值）的数据。

2. 数据预处理的方法

(1) 数据清洗——去噪声和无关数据；

(2) 数据集成——将多个数据源中的数据结合起来存放在一个一致的数据存储中；

(3) 数据变换——把原始数据转换成为适合数据挖掘的形式；

(4) 数据规约——主要方法包括数据立方体聚集、维度归约、数据压缩、数值归约、离散化和概念分层等。

3. 数据选取参考原则

(1) 尽可能赋予属性名和属性值明确的含义。

(2) 统一多数据源的属性编码。

(3) 去除唯一属性。

(4) 去除重复属性。

(5) 去除可忽略字段。

(6) 合理选择关联字段。

(7) 进一步处理：通过填补遗漏数据、消除异常数据、平滑噪声数据，以及纠正不一致数据，去掉数据中的噪音、填充空值、丢失值和处理不一致数据。

4. 数据预处理的知识要点

数据预处理相关的知识要点、能力要求和相关知识点见表3.1。

表3.1　数据预处理的知识要点

知识要点	能力要求	相关知识点
数据预处理的原因	(1) 了解原始数据存在的主要问题 (2) 明白数据预处理的作用和工作任务	(1) 数据的一致性问题 (2) 数据的噪音问题 (3) 原始数据的不完整和高维度问题
数据预处理的方法	(1) 掌握数据清洗的主要任务和常用方法 (2) 掌握数据集成的主要任务和常用方法 (3) 掌握数据变换的主要任务和常用方法 (4) 掌握数据规约的主要任务和常用方法	(1) 数据清洗 (2) 数据集成 (3) 数据变换 (4) 数据规约

5. 数据清洗的过程

(1) 读取数据。

(2) 和数据提供者讨论咨询。

(3) 数据分析(借助可视化工具发现脏数据)。

(4) 清洗脏数据(借助Matlab或者Java/C++语言)。

(5) 再次统计分析(最大值、最小值、中位数、平均值、方差等以及散点图)。

(6) 再次发现脏数据或者与实验无关的数据(去除)。

(7) 最后实验分析。

(8) 社会实例验证。

3.4.5 数据处理的基本流程与关键技术

1. 数据处理的整体框架

数据处理主要包括四个模块：分词（Words Analyze）、排重（Content Deduplicate）、整合（Integrate）和数据。

这四个模块的主要功能如下：

- 分词——对抓取到的网页内容进行切词处理。
- 排重——对众多的网页内容进行排重。
- 整合——对不同来源的数据内容进行格式上的整合。
- 数据——包含两方面的数据，Spider Data（爬虫从网页中抽取出来的数据）和 Dp Data（在整个数据处理过程中产生的数据）。

2. 数据处理的基本流程

整个数据处理过程的基本步骤如下：

（1）对抓取来的网页内容进行分词；

（2）将分词处理的结果写入数据库；

（3）对抓取来的网页内容进行排重；

（4）将排重处理后的数据写入数据库；

（5）根据之前的处理结果，对数据进行整合；

（6）将整合后的结果写入数据库。

3. 数据处理的关键技术——排重

排重就是排除掉与主题相重复项的过程，网页排重就是通过两个网页之间的相似度来排除重复项。Simhash 算法是一种高效的海量文本排重算法，相比于余弦角、欧式距离、Jaccard 相似系数等算法，Simhash 避免了对文本两两进行相似度比较的复杂方式，从而大大提高了效率。

采用 Simhash 算法来进行抓取网页内容的排重，可以容纳更大的数据量，提供更快的数据处理速度，实现大数据的快速处理。图 3.5 是 Simhash 的算法思路。

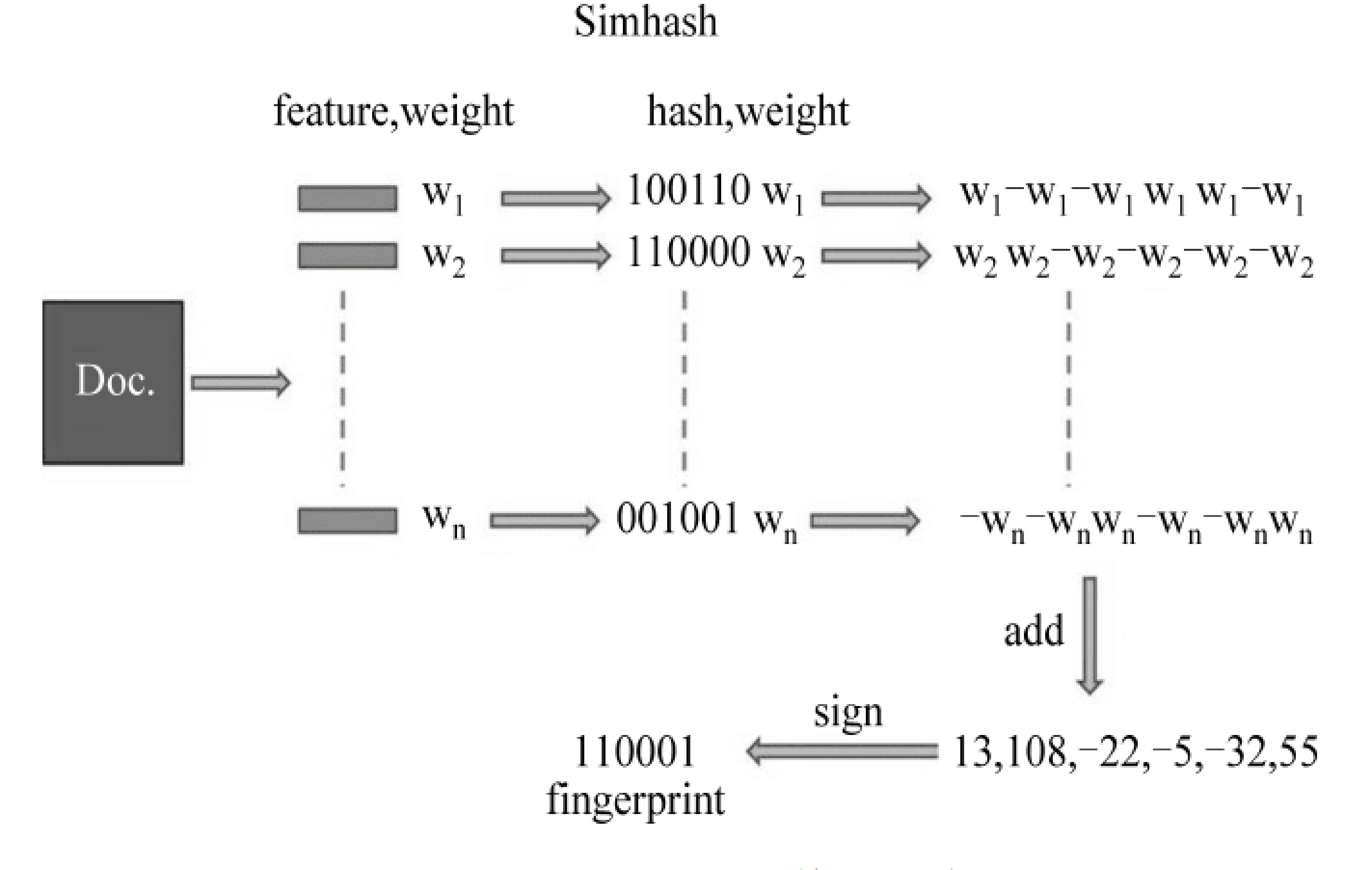


图 3.5 Simhash 的算法思路

Simhash 算法的基本思想描述如下：

输入为一个 N 维向量 V，比如文本的特征向量，每个特征具有一定权重。输出是一个 C 位的二进制签名 S。

(1) 初始化一个 C 维向量 Q 为 0，C 位的二进制签名 S 为 0。

(2) 对向量 V 中的每一个特征，使用传统的 Hash 算法计算出一个 C 位的散列值 H。对 $1 \leqslant i \leqslant C$，如果 H 的第 i 位为 1，则 Q 的第 i 个元素加上该特征的权重；否则，Q 的第 i 个元素减去该特征的权重。

(3) 如果 Q 的第 i 个元素大于 0，则 S 的第 i 位为 1；否则为 0。

(4) 返回签名 S。

对每篇文档根据 SimHtash 算出签名后，再计算两个签名的海明距离(两个二进制异或后 1 的个数)即可。根据经验值，对 64 位的 SimHash，海明距离在 3 以内的可以认为相似度比较高。

4. 数据处理的关键技术——整合

整合就是把抓取来的网页内容与各个公司之间建立对应关系。对于每一个公司来说，可以用一组关键词来对该公司进行描述，同样的，经过 DP 处理之后的网页内容，也可以用一组关键词来进行描述。因此，整合就变成了两组关键词(公司关键词和内容关键词)之间的匹配。

对于网页内容的分词结果来说，存在着两个特点：

(1) 分词结果的数量很大；

(2) 大多数的分词对描述该网页内容来说是没有贡献的。

因此，对网页的分词结果进行一下简化，使用词频最高的若干个词汇来描述该网页内容。

经过简化之后，两组关键词的匹配效率就得到了很大的提升，同时准确度也得到了保障；经过整合之后，抓取来的网页内容与公司之间就建立了一个对应关系，就能知道某个具体的公司有着怎样的数据了。

3.5　数据集成

数据集成的目的是运用一定的技术手段将各个独立系统中的数据按一定规则组织成为一个整体，使得其他系统或者用户能够有效地对数据进行访问。数据集成是现有企业应用集成解决方案中最普遍的一种形式。数据处于各种应用系统的中心，大部分的传统应用都是以数据驱动的方式进行开发。之所以进行数据集成，是因为数据分散在众多具有不同格式和接口的系统中，系统之间互不关联，所包含的不同内容之间互不相通。因此需要一种能够轻松访问特定异构数据库数据的能力。

3.5.1　数据集成的概念

数据集成是指将不同应用系统、不同数据形式，在原应用系统不做任何改变的条件

下，进行数据采集、转换和存储的数据整合过程。

3.5.2 数据集成面临问题

在信息系统建设过程中，由于受各子业务系统建设中具体业务要求和实施本业务管理系统的阶段性、技术性以及其他经济和人为因素等因素影响，导致在发展过程中积累了大量采用不同存储方式的业务数据。包括所采用的数据管理系统也大不相同，从简单的文件数据库到复杂的关系型数据库，它们构成了企业的异构数据源。异构数据源集成是数据库领域的经典问题，在构建异构数据源集成系统时，主要会面对以下几方面的问题。

1. 异构性

异构性是异构数据集成必须面临的首要问题，其主要表现在两个方面。

1）系统异构

数据源所依赖的应用系统、数据库管理系统乃至操作系统之间的不同构成了系统异构。

2）模式异构

数据源在存储模式上的不同。一般的存储模式包括关系模式、对象模式、对象关系模式和文档模式等几种，其中关系模式为主流存储模式。需要指出的是，即便是同一类存储模式，它们的模式结构可能也存在着差异。例如同为关系型数据库，Oracle 所采用的数据类型与 SQL Server 所采用的数据类型并不是完全一致的。

2. 完整性

1）异构数据

源数据集成的目的是为应用提供统一的访问支持。为了满足各种应用处理（包括发布）数据的条件，集成后的数据必须保证的完整性，包括数据完整性和数据集成的方法及技术。

2）数据集成

数据集成是指将不同应用系统、不同数据形式，在原应用系统不做任何改变的条件下，进行数据采集、转换和存储的数据整合过程。在企业数据集成领域，已经有了很多成熟的框架可以利用。目前通常采用基于中间件模型和数据仓库等方法来构造集成的系统，这些技术在不同的着重点和应用上解决数据共享和为企业提供决策支持。

面对以上几方面问题，产生了相关的数据变换技术和数据集成技术。

3.6 数据变换

自计算机诞生以来，人类积累了丰富的数据资源。计算机网络的普及，使得数据资源的共享成为一个热门话题。然而，由于时间和空间上的差异，人们使用的数据源各不相同，各信息系统的数据类型、数据访问方式等也都千差万别。这就导致各数据源、系统之间不能高效地进行数据交换与共享，成为“信息孤岛”。

用户在具体应用时,往往又需要将分散的数据按某种需要进行交换,以便了解整体情况。如,跨国公司的销售数据是分散存放在不同的子公司数据库中,为了解整个公司的销售情况,则需要将所有子系统的数据集中起来。为了满足一些特定需要,如数据仓库、数据挖掘等,也需要将分散的数据交换集中起来,以达到数据的统一和标准化。异构数据的交换问题由此产生,受到越来越多人的重视。

用户在进行数据交换时,面对的数据是千差万别的。产生数据差异的主要原因是数据的结构和语义上的冲突。异构数据不仅指不同的数据库系统之间的异构,如 Oracle 和 SQL Server 数据库,还包括不同结构数据之间的异构,如结构化的数据库数据和半结构化的数据。源数据可以是关系型的,也可以是对象型的,更可以是 Web 页面型和文本型的。因而,要解决数据交换问题,一个重要的问题就是如何消除这种差异。随着数据的大量产生,数据之间的结构和语义冲突问题更加严重,如何有效解决各种冲突问题是数据交换面临的一大挑战。

异构数据交换问题解决后,才会对其他诸如 OLAP、OLTP、数据仓库、数据挖掘、移动计算等提供数据基础。对一些应用,如数据仓库的建立,异构数据交换可以说是生死攸关。数据交换质量的好坏直接影响在交换后数据上其他应用能否有效进行。数据交换后,可以减小由于数据在存储位置上分布造成的数据存取开销;避免不同数据在结构和语义上差异造成的数据转换引起的错误;数据存放更为精简有效,避免存取不需要的数据;向用户提供一个统一的数据界面等。因此,数据交换对信息化管理的发展意义重大。

3.6.1　异构数据交换综述

异构数据交换技术的研究始于 20 世纪 70 年代中期,至今已有四十多年了。数据库的异构问题已经引起了各数据库厂家及许多数据库专家的注意。各数据库厂商积极参与国际标准的制定,他们新推出的产品都能支持统一的数据库语言、FAP 和 API 标准。其产品有的还留有支持新标准的余地,有的则采用了便于向国际标准过渡的形式。经过十几年对异构数据问题的探索和研究,人们已取得了不少成果,提出了许多解决异构数据交换的策略及方法,但就其本质可分成四类。

1. 使用软件工具进行转换

一般情况下,数据库管理系统都提供将外部文件中的数据转移到本身数据库表中的数据装入工具。比如 Oracle 提供的将外部文本文件中的数据转移到 Oracle 数据库表的数据装入工具 SQL Loader,Powersoft 公司的 PowerBuilder 中提供的数据管道(Data Pipeline)。

这些数据转移工具可以以多种灵活的方式进行数据转换,而且由于它们是数据库管理系统本身所附带的工具,执行速度快,不需要 ODBC 支持,在机器没有安装 ODBC 的情况下也可以方便地使用。

但是,使用这些数据转换工具的缺点是它们不是独立的软件产品,必须首先运行该数据库产品的前端程序才能运行相应的数据转换工具,通常需要几步才能完成,且多用手工方式进行转换。如果目的数据库不是数据转换工具所对应的数据库,数据转换工具就不

能再使用。

2. 利用中间数据库的转换

由于缺少工具软件的支持，在开发系统时可使用“中间数据库”的办法，即在实现两个具体数据库之间的转换时，依据关系定义、字段定义，从源数据库中读出数据通过中间数据库“灌”入到目的数据库中。

这种利用中间数据库的转换办法，所需转换模块少，且扩展性强；但缺点是在实现过程中比较复杂，转换质量不高，转换过程长。

3. 设置传送变量的转换

借助数据库应用程序开发工具与数据库连接的强大功能，通过设置源数据库与目的数据库两个不同的传送变量，同时连接两个数据库，实现异构数据库之间的直接转换。这种办法在现有的数据库系统下扩展比较容易，其转换速度和质量大大提高。

4. 通过开发数据库组件的转换

利用Java等数据库应用程序开发技术，通过源数据库与目的数据库组件来存取数据信息，实现异构数据库之间的直接转换。通过组件存取数据，关键是数据信息的类型问题，若源数据库与目的数据库对应的数据类型不相同，必须先进行类型的转化，然后双方才能进行赋值。

异构数据交换问题，实质上就是：一个应用的数据可能要重新构造，才能和另一个应用的数据结构匹配，然后被写进另一个数据库。它是数据集成的一个方面，也可以说是数据集成众多表现形式中的一种。

3.6.2 异构数据分析

异构数据交换的目标在于实现不同数据之间的数据信息资源、设备资源、人力资源的合并和共享。因此，分析异构数据，搞清楚异构数据的特点，把握住异构数据交换过程中的核心问题，是十分必要的。这样研究工作就可以做到有的放矢。

1. 异构数据

数据的异构性导致了应用对于数据交换的需求。那么何谓异构数据？异构数据是一个含义丰富的概念，它是指涉及同一类型但在处理方法上存在各种差异的数据，在内容上，不仅可以指不同的数据库系统之间的数据是异构的（如Oracle和SQL Server数据库中的数据）；而且可以指不同结构的数据之间的异构（如结构化的SQL Server数据库数据和半结构化的XML数据）。

总的来说，数据的异构性可以包括以下三个方面：系统异构、数据模型异构和逻辑异构。

1）系统异构

系统异构是指硬件平台、操作系统、并发控制、访问方式和通信能力等的不同，具体细分如下：

（1）计算机体系结构的不同，即数据可以分别存在于大型机、小型机、工作站、PC或

嵌入式系统中。

(2) 操作系统的不同，即数据的操作系统可以是 Microsoft Windows、Windows NT、各种版本的 UNIX、IBM OS/2、Macintosh 等。

(3) 开发语言的不同，比如 C、C++、Java、Delphi 等。

(4) 网络平台的不同，比如 Ethernet、FDDI、ATM、TCP/IP、IPX\SPX 等。

2) 数据模型异构

数据模型异构则是指 DBMS 本身的不同。比如数据交换系统可以采用同为关系数据库系统的 Oracle、SQL Server 等作为数据模型，也可以采用不同类型的数据库系统——关系、层次、网络、面向对象或函数型数据库等。

3) 逻辑异构

逻辑异构则包括命名异构、值异构、语义异构和模式异构等。比如语义的异构具体表现在相同的数据形式表示不同的语义，或者同一语义由不同形式的数据表示。

以上这些构成了数据的异构性，数据的异构给行业单位和部门等的信息化管理以及决策分析带来了极大的不便。因此异构数据交换是否迅速、快捷、可靠就成了行业、单位和部门制约信息化建设的一个瓶颈。

2. 冲突分类

异构数据之间进行数据交换的过程中，要想实现严格的等价交换是比较困难的。主要原因是由于异构数据模型间存在着结构和语义的各种冲突，这些冲突主要包括：

- 命名冲突——即源模型中的标识符可能是目的模型中的保留字，这时就需要重新命名。
- 格式冲突——同一种数据类型可能有不同的表示方法和语义差异，这时需要定义两种模型之间的变换函数。
- 结构冲突——如果两种数据库系统之间的数据定义模型不同，如分别为关系模型和层次模型，那么需要重新定义实体属性和联系，以防止属性或联系信息的丢失。

由于目前主要研究的是关系型数据模型间的数据交换问题，根据解决问题的需要，可将上述三大类冲突再次抽象划分为两大冲突：结构冲突和语义冲突。结构冲突是指需要交换的源数据和目标数据之间在数据项构成的结构上的差异。语义冲突是指属性在数据类型、单位、长度、精度等方面的冲突。对数据交换中需要解决的主要冲突，可做如下分类：

1) 结构冲突

结构冲突可分为两种情况：相似结构冲突和异构结构冲突。相似结构是指源和目标模式在表内部构成上相似，异构则与之相反。

(1) 相似结构冲突。

表相似结构冲突：如果两个表，表中的属性数量不同，但一个表的某些属性能够同另一个表某些属性对应，这时在这两个表之间产生了表结构冲突。

此时，两表在属性集上发生不一致性，表现为属性数量上的差异，但两表之间其他属性能够相互对应。其解决的方法一般为减少多余的属性或增加缺失的属性。

源和目标表中的属性之间存在以下两种情况：

① 源表的某些属性可以通过合并构成目标表的一个属性；

② 源表的一个属性经过分裂成为目标表的几个属性。

此时，源表和目标表产生了属性结构上的冲突。例如源表存在 Fname 和 Iname 两个属性，而目标表只有 Name 属性，但 Name 属性由 Fname 和 Iname 属性构成；则在源表的 Fname、Iname 属性和目标表的 Name 属性之间产生属性结构冲突。其解决的方法为在对应的冲突属性之间进行合并或分裂操作。

（2）异构结构冲突。

异构结构冲突可分为值-属性冲突、值-表冲突、属性-值冲突，表-值冲突等。以表 3.2 中几个表为例来说明表之间的异构结构冲突。

表 3.2　异构结构冲突示例

Dalian

Date	Number
31/10/03	10012
31/11/03	10091

Yantai

Date	Number
31/10/03	5983
31/11/03	9832

Qingdao

Date	Number
31/10/03	78934
31/11/03	78965

港口统计表(table_Value_port)

Date	Dalian	Yantai	Qingdao
31/10/03	10012	5983	78934
31/11/03	10091	9832	78965

总公司统计表(table_Value_company)

Date	Number	Company
31/10/03	10012	Dalian
31/11/03	10091	Dalian
31/10/03	5983	Shenyang
31/11/03	9832	Shenyang
31/10/03	78934	Qingdao
31/11/03	78965	Qingdao

其中 Dalian(大连)、Yantai(烟台)、Qingdao(青岛)三个表表示位于三地的子港务公司每月的集装箱出口数量表，表 Table_value_port 是港口集装箱出口统计表，而表 Table_value_company 是总公司的集装箱出口数量统计表，它是由 Dalian、Yantai、Qingdao。三个表中的数据经过数据交换后得到的。

属性-值冲突：如果相同的信息在一个表中被表示为属性的名称而在另一个表中被表示为属性的值时，则产生了属性-值冲突。

如总公司统计表(Table_Value_company)中 Company 属性的某个值(如 Dalian)在

利润表(Table_value_port)表中成为一个属性的名称。

表-值冲突：当数据库中表的某个属性值被表示为一个表的名字时，则产生了表-值冲突。如总公司统计表(Table_value_Companys)中 Company 属性的某个值(如 Yantai)成为表 Yantai 的名称。

对异构的情况，比较常见的转换为“表”到“值”的转换和“属性”到“值”的转换。

对“值”到“表”，“值”到“属性”，“属性”到“表”，“表”到“属性”的转换，由于实际数据交换中，目标系统表结构很少采用这种设计方式，因而研究重点是“表”到“值”、“属性”到“值”两种异构情况的转换。

2）语义冲突

语义冲突主要分为两种情况：表的语义冲突和属性语义冲突。表的语义冲突是指具有相同标识符的表语义不同。属性语义冲突是指属性的数据类型、单位、格式等的冲突。

(1) 表的语义冲突。

表的语义冲突是指具有相同或相似结构的两个表在语义上的差异。如一个表为所有员工的工资，而另一个结构相同的表则为某个部门员工的工资。对相同的结构，只需要将所有源表数据合并到目标表或将源表水平分割为各个目标表即可。

(2) 属性语义冲突。

数据类型冲突：同一属性的数据在不同表中的数据类型不一致。如年龄在一个表中为字符型而在另一个表中为数值型。其解决办法为将一种数据类型转化为另一种数据类型。

命名冲突：表示同一概念的属性在不同表中命名不一样。如，一个表中用 Company 属性表示公司，在另一个表中用 Corporation 属性表示公司，对应的属性在命名上有差异。解决的办法是统一属性的命名。

单位冲突：同一属性在不同表中，其值的单位不一样。如，一个表中身高以米为单位，另一个表中用厘米为单位。此时，对应属性在度量单位上有差异。解决办法是统一单位。

数据长度冲突：属性值的长度不一样。

数据精度冲突：同一属性的值在不同表中的数据精度不一样。如，一个表中工资值为 100.89，在另一个表中为 100.9。解决办法是进行精度转换。

数据格式冲突：同一属性的值在不同表中的表现格式不一样。最典型的例子如日期，一个表中为“MM/DD/YY”格式，在另一个表中为“YY/MM/DD”。此时，对应属性在数据格式上出现差异。解决的办法是统一数据的表现格式。

其他情况：这类情况比较特殊，如物理运动的测量是由于参照物选择不同引起的测量值的差异。可根据实际交换时的情况进行分析。

总之，在进行数据转换时，一方面源数据模式中所有需要共享的信息都转换到目标数据中，另一方面这种转换又不能包含冗余的关联信息。

3.6.3　异构数据交换方式

异构数据交换就是实现分布式网络环境下，不同位置、平台和格式的数据以一种统一

的交换标准集中展现给用户，并可以进行数据资源的抽取和利用。

异构数据存放于异构数据库中，异构数据库的各个组成部分具有自治性和数据库管理系统，实现数据共享的同时又保持自己的应用特性、完整性控制和安全性控制，确保基于异种系统平台实现对异构数据库的查询和联合使用。

提供一个独立于特定的数据库管理系统的统一编程界面。异构数据库系统是相关的多个数据库系统的集合，目标在于实现不同数据库之间的资源的合并和共享，为应用系统提供安全的、统一的、快捷的信息查询、数据挖掘和决策支持服务。

异构数据库系统的数据交换主要是为了消除异构数据之间的冲突，通过一些设备在不同的应用平台和操作系统之间使交换数据的双方可以实现彼此之间的透明访问和各系统间的数据共享、业务协同，从而解决了信息孤岛问题。

异构数据交换方式主要分为数据发布、数据集成和交易自动化。

1. 异构数据的发布

异构数据的发布指的是将异构数据库中的数据根据用户设定的条件及提取出来的目标信息，按照数据请求者要求的、可以接受的格式发送出去。

2. 异构数据的集成

异构数据的集成指的是根据用户设定的条件及提取出来的目标信息将异构数据源集成起来并且提供给用户一个统一的视图(物理的、逻辑的)。异构数据的集成屏蔽了数据源的异构性。可以使应用程序以统一的方式对不同分布的、结构异构的数据源进行访问，可以为这些数据源提供实时的读写操作，也可以完成各个业务模块之间的数据共享，从而畅通无阻地实现彼此之间的通信，进而理顺业务操作过程。

异构数据集成体系结构主要有三种：联邦数据库、Mediator/Wrapper 模式以及数据仓库。

1) 联邦数据库

联邦数据库系统是实现数据库集成问题的一种传统方法，是在任何两种异构数据源之间建立起彼此互相转化的方式。这种模式的数据集成是个 N 维问题，假设存在 N 个彼此异构的数据库系统，并且任意两个之间要实现彼此转换，则需要实现的转换模式总和为 $T=N(N-1)$。因此，使用这种方式时，开发人员要编写 $N(N-1)$段代码来实现两两之间的彼此共享。

2) Mediator/Wrapper 模式

Mediator/Wrapper 模式是一种软件构件。通过为所有异构数据源提供一个统一的虚拟视图的方式来实现集成目。这种集成方式并不需要存储任何实际数据，只需要系统为用户提供一个全局模式(即 Mediator 模式)，用户只需要针对全局模式提交查询条件，而不需要知道数据源的模式、位置以及访问方法，系统会自动地将用户的查询条件分别转换成一个或多个对数据源的查询，再将查询得到的结果集进行处理和整合，最终返回给用户。

Mediator/Wrapper 模式中的异构数据源具有完全的自治性，从而可以方便地对数据源进行添加和删除。中介系统一般由一个 Mediator 和多个 Wrapper 构成，Mediator 的作用是

将针对全局模式的查询进行分析，然后分解成若干个子查询，并将它们分别转换成针对所对应数据源的查询，最后将所有数据源的结果进行合并和整合，再返回给用户。Wrapper 的作用是将各个数据源中的数据转换为统一集成系统可以处理的结构化的数据。

Mediator/Wrapper 这种方式的优点是可以实现大量的数据源的互访和通信，对数据源的数目并没有限制，但是系统的结构和内部处理算法实现起来十分复杂。

3）数据仓库

数据仓库集成异构数据源的策略是将来自几个异构数据源的数据副本，按照一个集中、统一的视图要求，进行预处理、转换，以符合数据仓库的模式，并存储到数据仓库中。这样，对于使用者来说感觉就像在使用一个普通的数据库一样。

一旦数据存储于数据仓库，用户使用查询就像是在原来单一的数据源中查询一样。另一方面，数据仓库可能会禁止用户去更新数据，因为，用户对数据仓库中数据的更新将不会反映到原来的数据源中，这就会造成数据源和数据仓库中数据不一致的问题。

目前，进行数据仓库中数据构建的方式有以下三种：

（1）数据仓库周期性的从原数据源中重新构建数据。

最常使用的方式是在每天午夜（那时系统可能需要关机，并且不是用户使用数据仓库的高峰期）或者是更长周期的午夜时刻进行数据重建。这种方式的主要缺陷是需要将数据仓库关闭，而事实上数据的重建可能需要很长的时间。对于某些应用来说，过长的时间会使很多数据过时。

（2）数据仓库周期性地从原数据源中更新数据（采用增量更新的模式，即每次数据仓库更新上次更新以后修改的数据）。

这种方式只会影响到数据仓库中少量的数据，这样即使是在数据仓库的容量很大的时候，数据更新的时间也不会很久。该方式主要的缺点是用于计算数据仓库中数据更新的算法（增量更新算法），相对于从原始数据开始构建数据仓库的算法要复杂得多。

（3）数据仓库即时更新异构数据源的数据变化。

当一个或多个数据源中的数据发生变化的时候，立即更新数据仓库中相应的数据。由于这种方法需要数据仓库和数据源之间频繁的通信，所以这种方式只适用于小型的、数据更新量小的数据仓库中。这种方式有着一个典型而且广泛的应用——自动股票交易系统。

总之，数据仓库模式的异构数据库数据共享集成的优点是便于进行联机分析和数据挖掘；缺点是数据重复存储，难以及时更新。综上所述，三种集成方式各有优缺点，我们应该根据实际应用的具体要求和特点来选择最适合的集成方式以满足具体应用的实际要求。

3. 交易自动化

各种应用只要遵循共同的标准，就可以使得应用程序开发商开发出具有一定自动处理能力的代理程序，从而提高工作效率。

3.6.4　异构数据交换技术

实现异构数据交换的方法和技术较多，这里列出 XML、本体技术、Web Service 等几项技术。

1. 基于 XML 的异构数据交换技术

XML(Extensible Markup Language,可扩展标记语言)是 SGML(Standard Generalized Markup Language,标准通用标记语言)的一个简化子集,1998 年 2 月成为 W3C(The World Wide Web Consortium 互联网联合组织)标准。

XML 提供了一种灵活的数据描述方式。XML 支持数据模式、数据内容、数据显示方式三者的分离的特点,这使得同一数据内容在不同终端设备上的个性化数据表现形式成为可能,在数据描述方式上可以更加灵活。XML 具有很强的链接能力可以定义双向链接、多目标链接、扩展链接和两个文档间的链接。

XML 具有自描述性。XML 文档通常由模式描述文件和事例文件组成,前者用于描述 XML 事例文件所能使用的标记、标记的结构、标记的含义等,而 XML 事例文件则使用这些预定义的标记描述数据,所以 XML 具有自描述性。

XML 简单,易于处理。从数据处理的角度看,XML 足够简单,易于阅读,又易于被应用程序处理。

上述特点使得 XML 可以为结构化数据、半结构化数据、关系数据库、对象数据库等多种数据源的数据内容加入标记,适于作为一种统一的数据描述工具,扮演异构应用间数据交换载体或多源异构数据集成全局模式的角色。事实上,XML 已经成为 Internet 环境下数据表达的公开而被广泛支持的标准。

1) 基于 XML 的异构数据交换的总体过程

由于系统的异构性,需要交换的数据具有多个数据源,不同数据源的数据模式可能不同,导致源数据和目标数据在结构上存在差异。

在进行数据交换时,首先必须将数据模型以统一的 XML 格式来描述,这就需要使用 XML 的 DTD 或 XML Schema 来定义文档的结构,DTD 定义 XML 文档的基本结构,但不涉及任何有关的实际数据,通过定义适当的 DTD 将源数据库中的数据转换成 XML 文档,然后使用 DOM 技术来解析 XML 文档,这样就可以将 XML 文档中的数据存入目标数据库,从而实现了异构数据的交换。

由于 DTD 文档定义的数据结构与源数据库中的数据结构保持一致,从而保证了生成的 XML 文档与源数据库中数据的一致性。

其总体交换过程如图 3.6 所示。

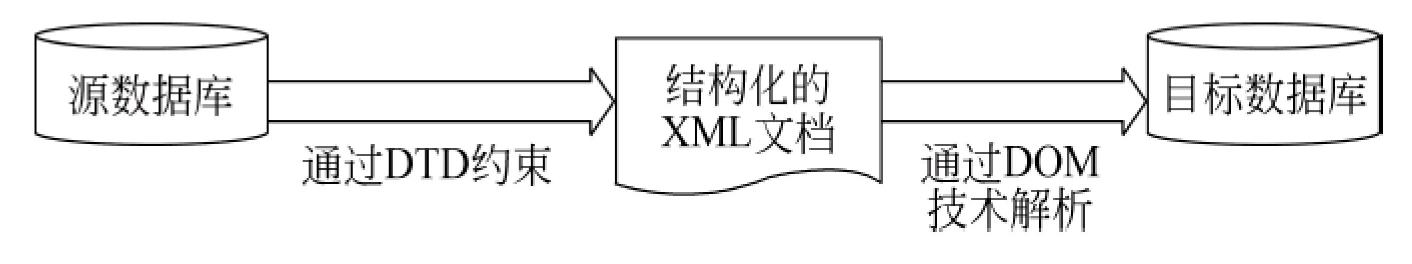


图 3.6 基于 XML 的异构数据交换的总体过程

2) 数据库数据与 XML 文档的映射原理

在 XML 数据和数据库之间转换时,需要考虑许多问题,XML 不支持任何有实际意义的数据模型,所有 XML 文档中的数据都会被当成纯文本处理。通常数据转换中间件需要把 XML 文档中的纯文本转换成数据库的数据类型,或把数据库的数据类型转换为纯文本的 XML 格式。在 XML 文档结构和数据库模式结构之间进行相互映射,一般有两

种映射方法：模板驱动映射与模型驱动映射。

（1）模板驱动映射。

基于模板驱动的映射是一种浅层次的映射，是一种基于模板的DTD到关系模式的转换算法，其转换比较简单，只要给出模板，就可以快速生成相应的XML文档。基于模板的映射方法不用预定义XML数据与数据库数据之间的映射关系，只是在XML文档中嵌入带参数的SQL命令，这些模板中的命令由数据转换中间件来处理，在转换过程中被识别和执行，将执行的结果替换到命令所在的位置上，从而生成XML文档。因为使用模板驱动映射在数据转换时需要生成大量合理的模板，所以系统要为用户提供生成模板的工具，以及相应的指令执行程序，其过程如图3.7所示。

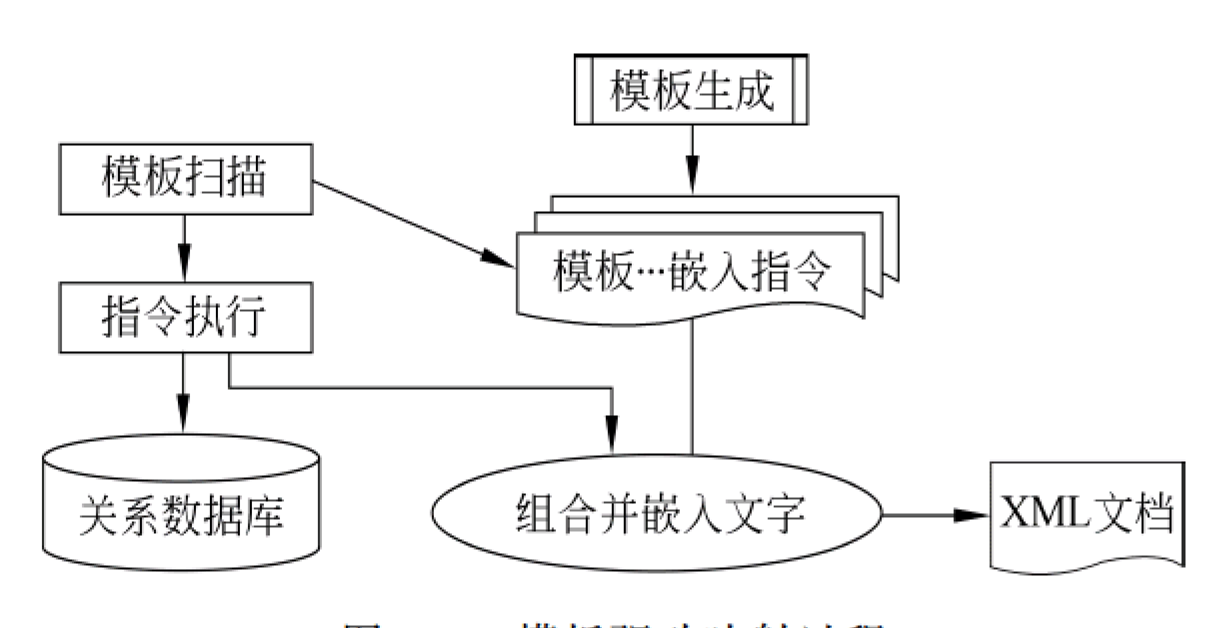


图3.7 模板驱动映射过程

基于模板映射的优点是转换步骤简单，查询语言灵活性大，支持通过HTTP的传递参数，允许嵌套查询，支持SELECT语句的参数化，支持编程结构，如可以由程序构建loop循环或if判断等。目前大多数的数据库产品都属于模板映射，如SQL Server、DB2和Oracle等。缺点是模板驱动映射是以XML内嵌的SQL执行的数据结果集为依据，不涉及数据库赖以存在的数据模型，只能将关系数据库的数据转换为XML文档，并舍弃了关系模式的约束条件，所以也不支持反向的转换。

（2）模型驱动映射。

模型驱动映射是一种深层次的映射，其原理是利用XML文档中的数据模型的结构显性或隐性地映射成其他数据模型的结构。实现数据库和XML文档间的数据转换的关键是在数据库模式和XMLSchemas或DTD之间建立映射关系，用具体的模型来实现数据间的映射。通常关系数据库利用关系型，面向对象数据库利用对象模型，而XML文档依赖的是Schemas或DTD。当数据从数据库转换成XML文档时，因为依照的是单个模型，通常需要结合XSL来控制模板驱动，从而保证了系统的灵活性。

要实现关系数据库数据转换XML文档时，将层次结构的XML文档理解成一张二维表，直接与数据库中的关系表相对应，把表或查询结果的数据插入到XML文档的相应位置便可，相反把XML文档数据转换成数据库数据时，只要把内容插入到相应的二维表中即可。如果是把对象数据库中的数据转换为XML文档时，首先要将XML文档映射成同样具有层次结构的对象树(DOM)，然后将对象树映射到面向对象的数据库中，或通过“对象-关系技术”将对象树映射到关系数据库中，其过程如图3.8所示。

基于模型映射转换的优点是有数据模型的支持，相对比较简单，可以实现XML数据与数据库数据间的双向映射。缺点是XML文档结构受数据模型的限制，不够灵活，不适

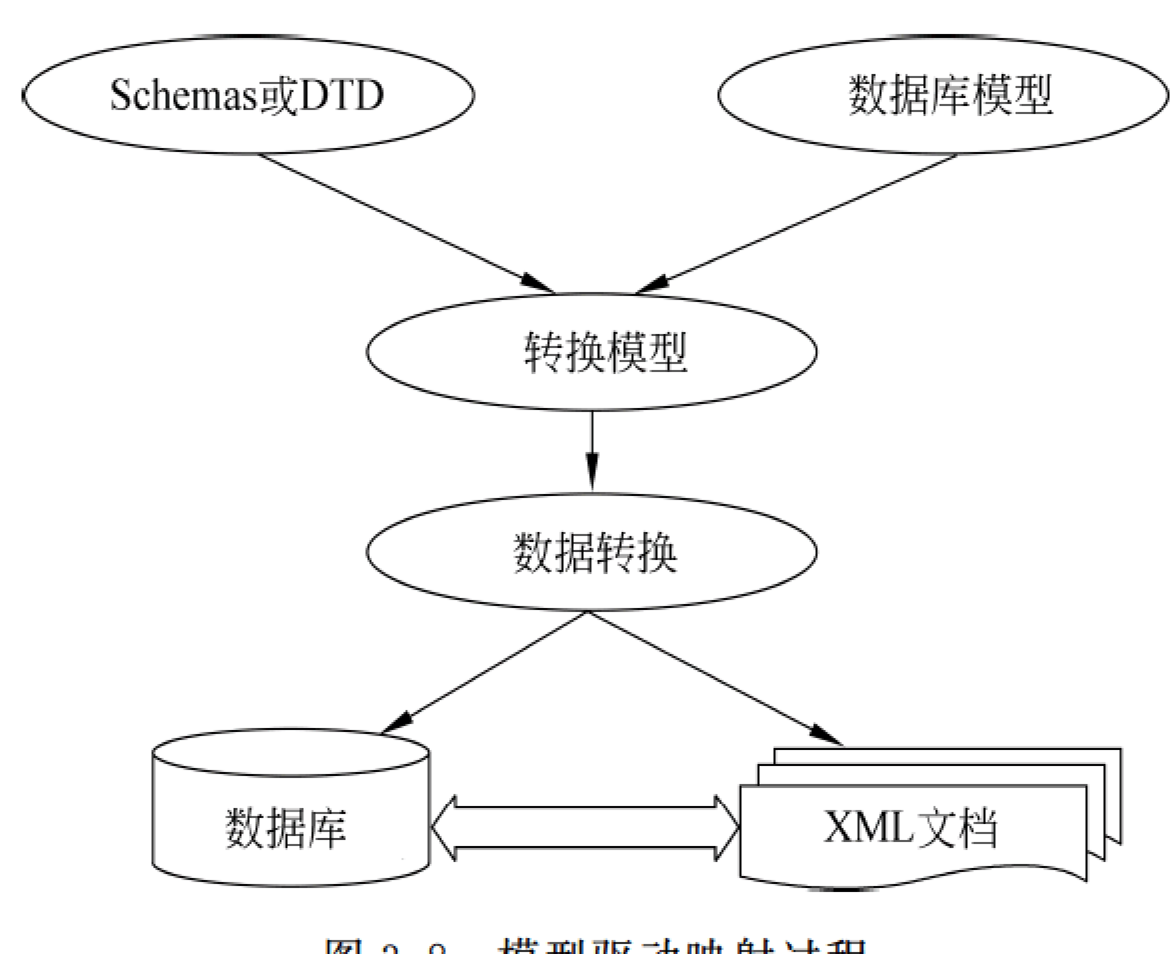


图 3.8　模型驱动映射过程

用于嵌套层次比较深的 XML 文档进行映射，也不能适用于多个对象集合的映射，映射的时候表的结构必须与对象结构一致，对结构不一致的数据表也很难映射，不能定制数据库数据与 XML 的映射。

2. 本体技术

本体是对某一领域中的概念及其之间关系的显式描述。是语义网络的一项关键技术。本体技术能够明确表示数据的语义以及支持基于描述逻辑的自动推理。为语义异构性问题的解决提供了新的思路，对异构数据集成来说应该有很大的意义。

但本体技术也存在一定的问题：已有关于本体技术研究都没有充分关注如何利用本体提高数据集成过程和系统维护的自动化程度、降低集成成本、简化人工工作。基于语义进行自动的集成尚处于探索阶段，本体技术还没有真正发挥应有的作用。

3. Web Service 技术

Web Service 是近年来备受关注的一种分布式计算技术。它是在 Internet 或 Intranet 上使用标准的 XML 语言和信息格式的全新的技术架构。其内容主要包括 WSDL(Web Service 描述语言，用于进行服务描述)、UDDI(统一描述、发现和集成规范，用户服务的发布和集成)和 SOAP(简单对象访问协议，用于消息传输)。

从用户角度看，Web Service 就是一个应用程序，它向外界暴露出一个能够通过 Web 进行调用的 API。服务请求者能够用非常简便的类似于函数调用的方法通过 Web 来获得远程服务，服务请求者与服务提供者之间的通信遵循 SOAP 协议。

Web Service 体系结构由角色和操作组成。角色主要有服务提供者(Service Provider)、服务请求者(Service Requestor)、服务注册中心(Service Registry)。操作主要有发布(Publish)、查找(Find)、绑定(Bind)、服务(Service)、服务描述(Service Description)，其具体架构如图 3.9 所示。

其中，“发布”是为了让用户或其他服务知道某个 Web Service 的存在和相关信息，“查找”是为了找到合适的 Web Service，“绑定”则是在提供者与请求者之间建立某种联系。

在异构数据库集成系统中，可以利用 Web Service 具有的跨平台、完好封装及松散耦

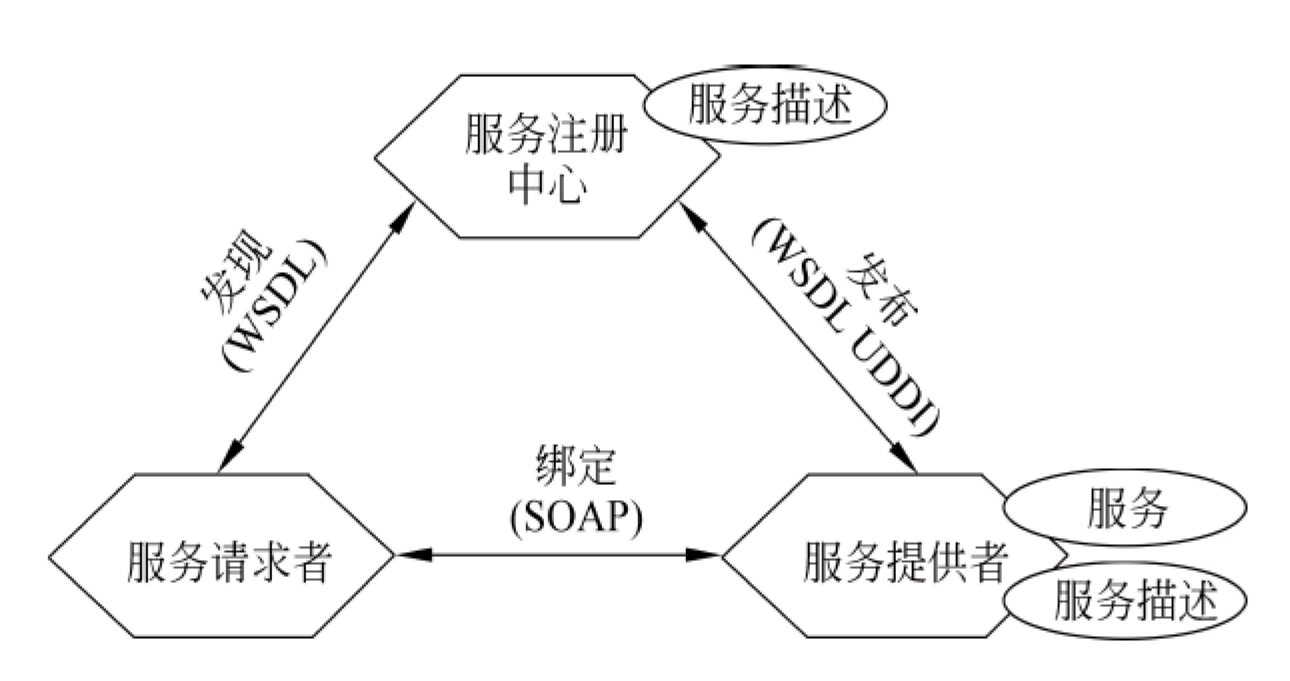


图 3.9　Web Service 架构

合等特性，对每个数据源都为其创建一个 Web Service，使用 WSDL 向服务中心注册，然后集成系统就可以向注册中心发送查找请求并选择合适的数据源，并通过 SOAP 协议从这些数据源获取数据。这样不仅有利于数据集成中系统异构问题的解决，同时也使得数据源的添加和删除变得更加灵活，从而使系统具有松耦合、易于扩展的良好特性，能实现异构数据库的无缝集成。

3.6.5　异构数据交换与集成的研究方向

鉴于异构数据交换所固有的特点，可以相信，异构数据交换会随着各个难题的解决而得到越来越广泛的应用。

今后，异构数据交换与集成的研究方向应该包括：

(1) 基于网格、本体语义的数据集成方案的研究。

(2) 集成数据的完整性、一致性约束。

(3) 半结构化数据全局模式的构建方法和映射方法。同样要保证数据的完整性和一致性约束能够在半结构化的数据间传递。

(4) 数据集成过程中安全、可靠的数据传输技术。

3.7　大数据应用案例之：互联网行业哪个职位比较有前途

互联网行业的迅猛发展，使得越来越多的年轻人投入到互联网的浪潮中。互联网公司需求哪些人才，哪一类职业更抢手，哪些人更容易在互联网公司找到工作，各类职业工作年限对应年薪分布如何，哪些城市互联网公司发展得更好，各个细分领域的互联网公司对人才的需求如何？下面就用数据的方式来对互联网行业的职场进行分析。

1. 数据来源

数据来源于专注互联网招聘的垂直领域网站——拉勾网，采集时间：2014.9—2015.9，涉及 756 000 个发布职位。

本报告使用了超过 75 万个独立的真实发布职位，100 000 家互联网公司，职位来自 10 万家互联网公司，266 个不同城市区域。

2. 互联网各类职位需求状况

整个互联网行业是建立在计算机技术开发的基础之上，因此该行业对于技术类人才

的需求占了45%左右。然而现在的互联网产品模仿非常严重，新产品上线不久往往就有很多的竞争者，加之现在的互联网产品中技术越来越不能成为其壁垒，那么，除了产品自身优秀外，市场和运营的作用就非常关键，可以说决定着产品的前途和命运。

从图3.10可以看到，互联网行业对于市场和运营的人才需求比例也非常大。从排在前三类职位的细分职业来看，互联网行业对研发工程师、销售人员、运营专员的需求分别占了各自所属类别职位的一半以上。

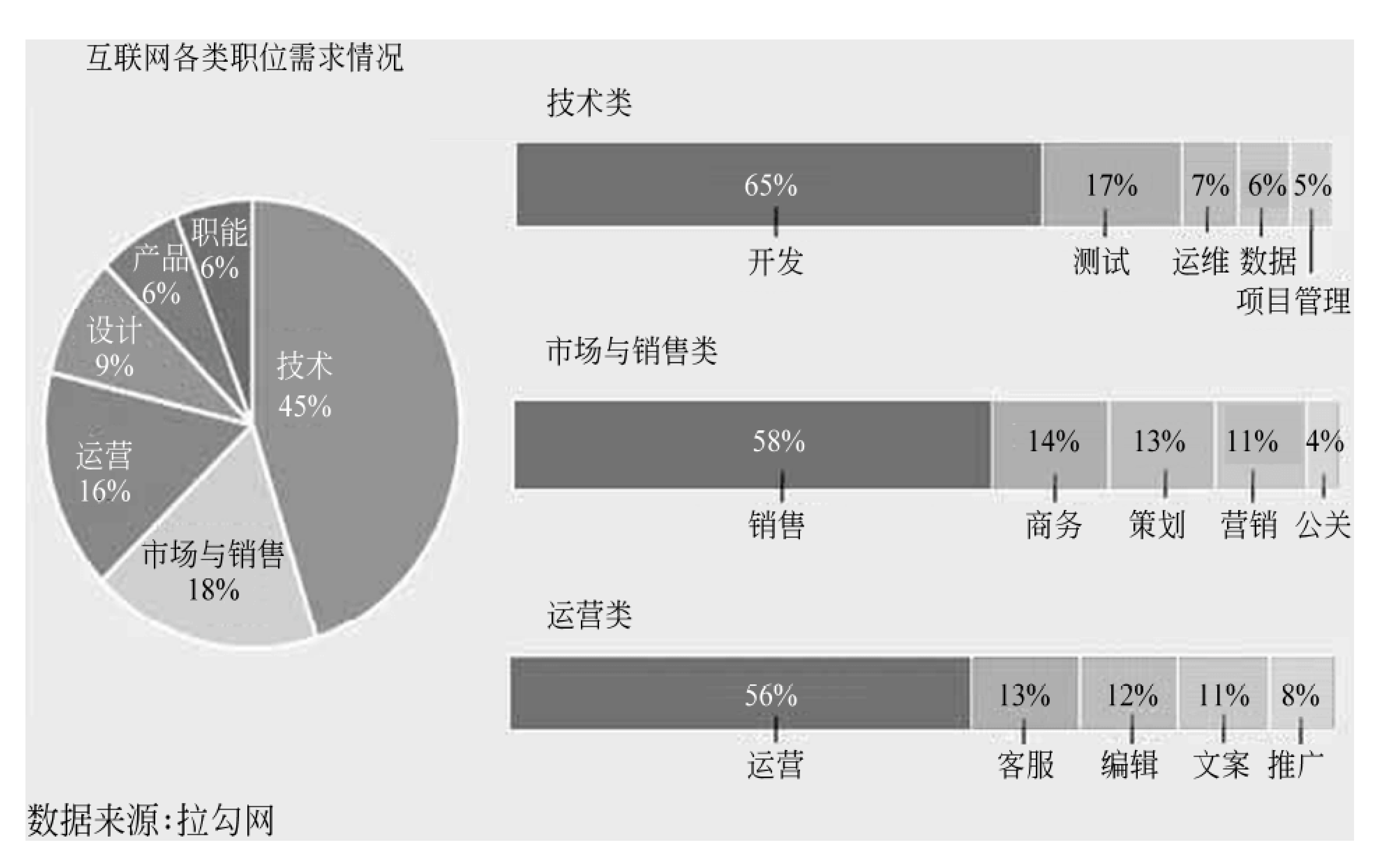


图3.10 互联网各类职位需求状况

3. 互联网最难招/易招职位

根据职位从开放到关闭时所经历的平均天数来衡量各个职位的难易招程度。从图3.11可以看到，互联网公司招聘一名营销人员平均需要54.4天时间，可谓互联网最难招的职位，排名前5的最难招职位中，有2个职位都属于市场与销售类别，这应该是和目前互联网大量面向客户项目的创立，对市场与销售人员的庞大需求量成正相关，同时由于



图3.11 互联网最难招/易招职位

互联网市场类职位的起薪相对较低，也成为该类职位难招到人的制约因素。

我们看到，互联网最易招的5种职位中，有4种均属于职能类别的职位，表明互联网对这类职位人员的需求量不大。我们发现前段时间兴起的新兴职位：程序员鼓励师属于互联网最易招的职位，一方面是目前行业内公司对该职位需求量较小，要求不高；另一方面由于其有趣的工作职责要求，吸引了很多年轻女性前来应聘。

4. 互联网5大抢手职业

定义一个职业的抢手程度＝平均月薪×发布职位数/已招到职位数，根据这个公式，我们统计出排名前5的互联网抢手职业，如图3.12所示，可以看到，技术岗位职业占据了4席，架构师由于对于其高要求的技术能力需求成为最抢手的职业，产品经理也属于5大抢手职业之一，这对于那些不需要特别精通技术，又想在互联网行业发展的朋友无疑是一个很好的消息。

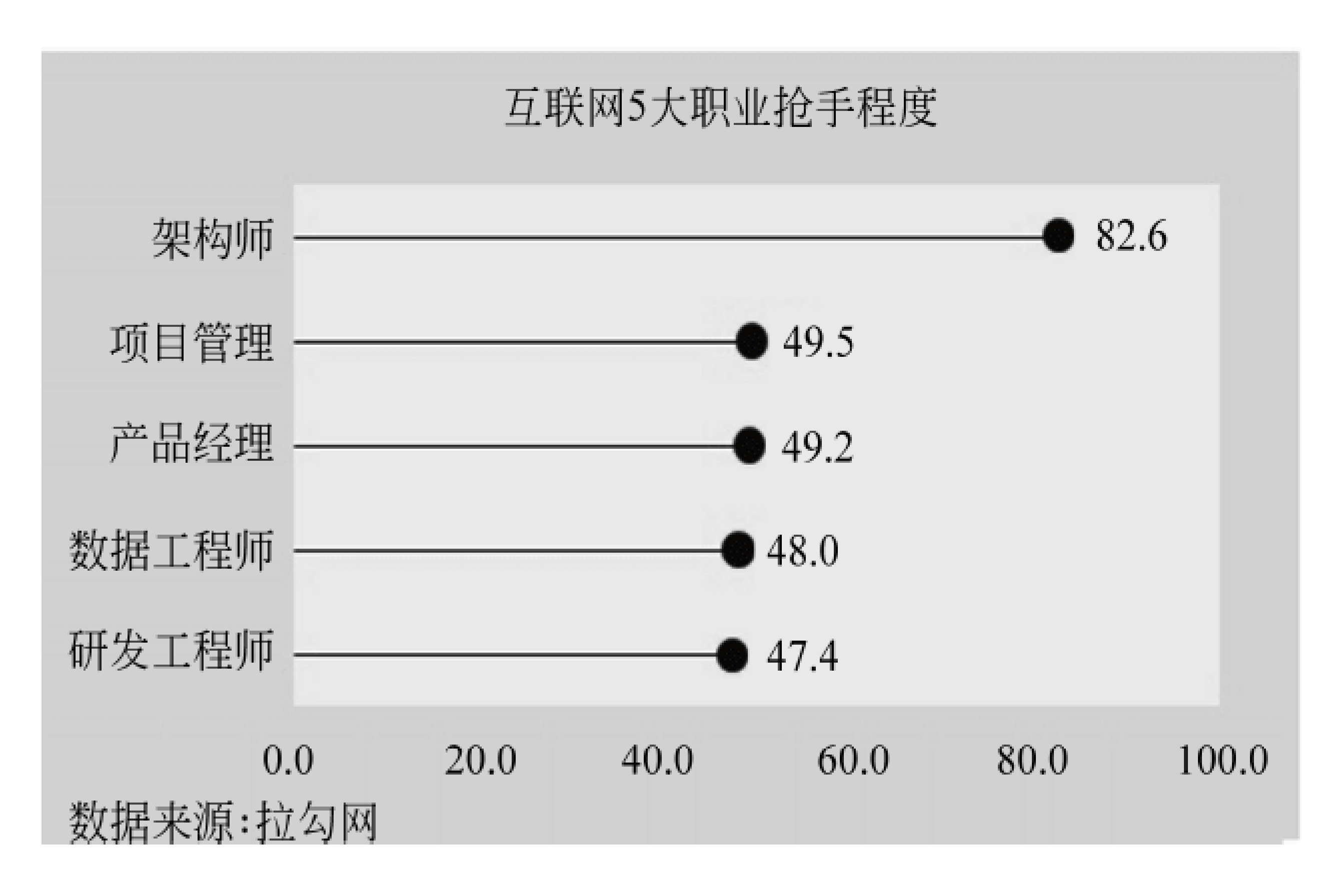


图3.12　互联网5大抢手职业

5. 互联网5大过剩职业

与抢手职业计算公式相同，统计出得分最低的5个职业，从图3.13可以看到，这些职业均属于职能类别，由于很多互联网公司属于初创期，对于财务方面的业务往往不重视，要么外包给财务公司，要么某个人员兼任，所以出纳这个职业成为互联网行业最过剩的职业。

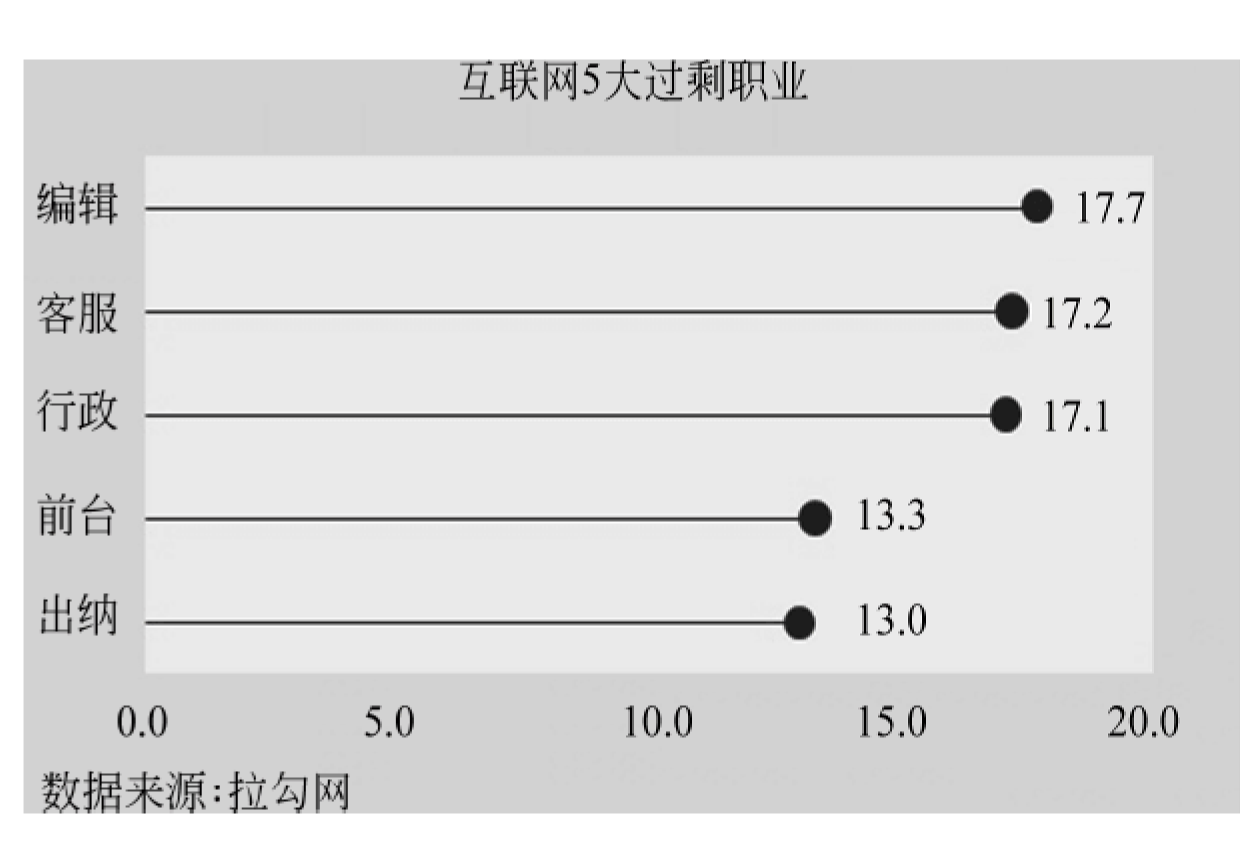


图3.13　互联网5大过剩职业

6. 互联网工作年限与对应年薪

从图 3.14 可以看到，前 5 年里，技术和产品类别的职位年薪属于互联网行业中较高的群体，工作 5 年后，运营类别的职位年薪有了较大的涨幅，后期甚至超过了做产品的人员。

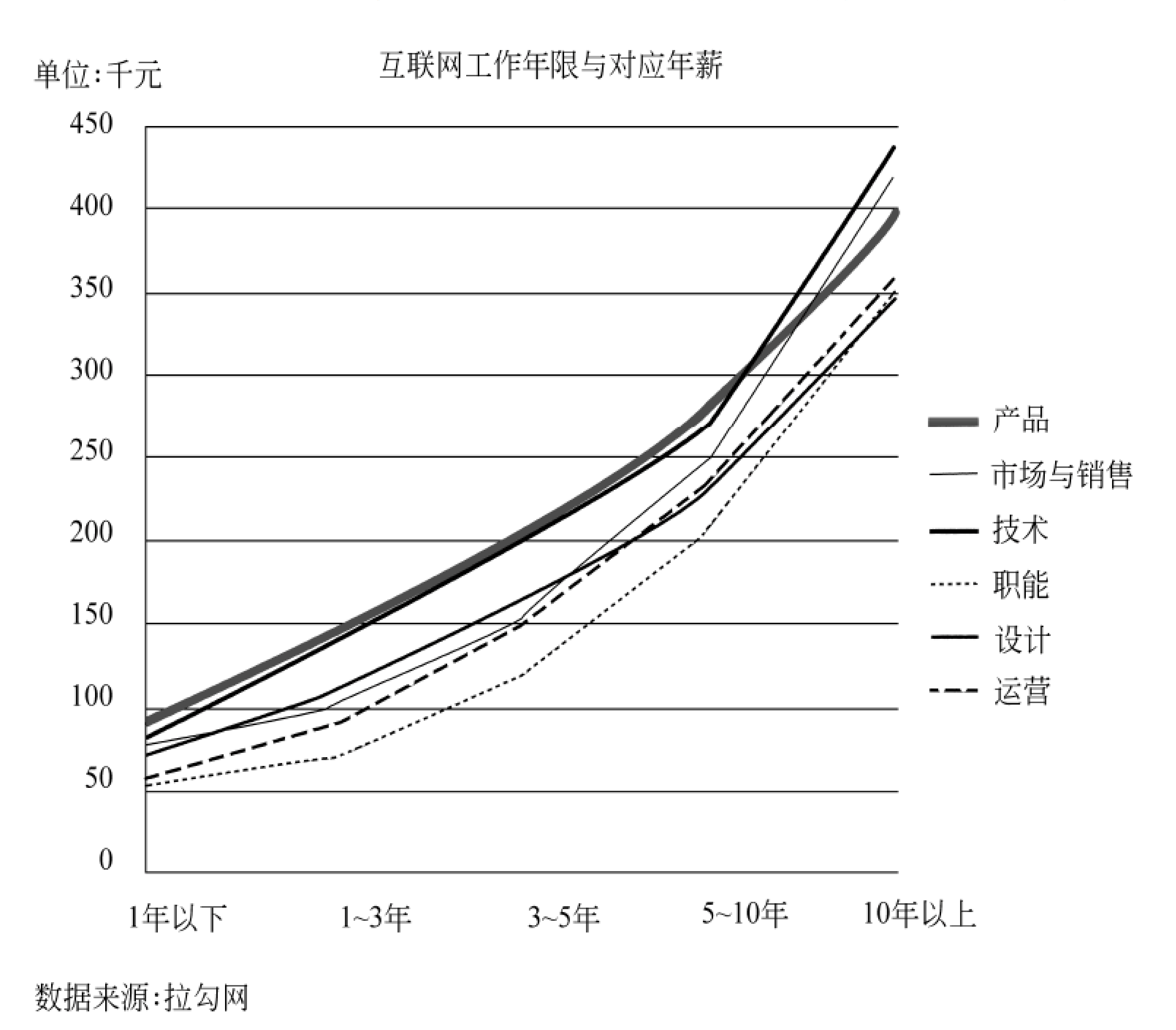


图 3.14 互联网工作年限与对应年薪

7. 各个城市互联网公司发展状况

选取互联网公司最集中，排名前 5 的城市，从图 3.15 可以看到，上海的非天使轮公司

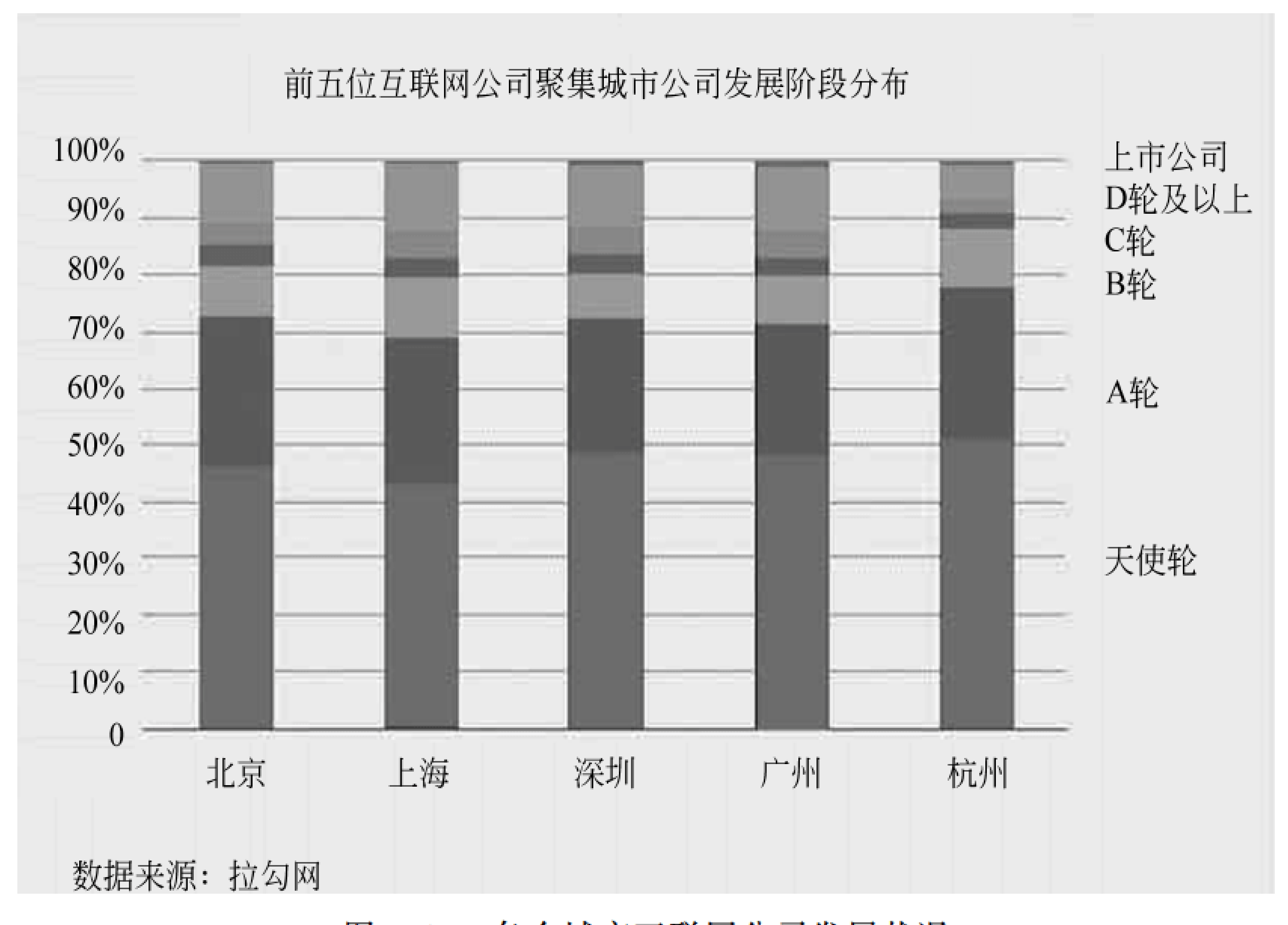


图 3.15 各个城市互联网公司发展状况

占比最多，上市公司占比也最高，表明上海的创业公司发展还不错，准备创业的人可以考虑以上海作为创业地。

8. 互联网细分行业统计

根据互联网公司的细分行业，对每个行业互联网公司的每日平均岗位数、平均月薪、平均公司规模进行了统计，从图 3.16 可以看到，移动互联网、搜索、大数据和游戏行业的公司发展都不错，薪酬待遇相应也属于行业的前列。

互联网细分行业统计

公司所属行业	平均岗位数(每日)	平均薪水(月薪/元)	平均公司规模(人)
1 O2O	210	9939	206
2 云计算大数据	151	10424	340
3 企业服务	174	9132	316
4 健康医疗	66	9267	159
5 在线旅游	47	10976	339
6 媒体	75	9162	276
7 招聘	44	12125	212
8 搜索	40	17348	1446
9 教育	105	9363	243
10 智能家居	27	8694	155
11 游戏	156	11293	245
12 生活服务	65	10052	441
13 电子商务	506	9551	281
14 硬件	37	9805	228
15 社交	66	10695	86
16 移动互联网	1280	10565	289
17 运动体育	15	10328	485
18 金融互联网	242	10538	307

数据来源:拉勾网

图 3.16　互联网细分行业统计

习题与思考题

一、选择题

1. 下面哪种不属于数据预处理的方法？（　　）
   A. 变量代换　　B. 离散化　　C. 聚集　　D. 估计遗漏值
2. （　　）的目的缩小数据的取值范围，使其更适合于数据挖掘算法的需要，并且能

够得到和原始数据相同的分析结果。

A. 数据清洗　B. 数据集成　C. 数据变换　D. 数据归约

3. Google 收集的信息不包括(　　)。

A. 日志信息　B. 位置信息

C. 你的家庭成员　D. Cookie 和匿名标识符

4. 大数据的取舍与(　　)不相关。

A. 易于提取　B. 家庭信息

C. 数字化　D. 廉价的存储器

5. 大数据,或称巨量资料,指的是所涉及的资料量规模巨大到无法透过目前主流软件工具,在合理时间内达到撷取、管理、处理、并(　　)成为帮助企业经营决策更积极目的的信息。

A. 收集　B. 整理　C. 规划　D. 聚集

6. 下面哪种不属于数据预处理的方法?(　　)

A. 变量代换　B. 离散化　C. 聚集　D. 估计遗漏值

7. 数据清洗的方法不包括(　　)。

A. 缺失值处理　B. 噪声数据清除

C. 一致性检查　D. 重复数据记录处理

8. 智能健康手环的应用开发,体现了(　　)的数据采集技术的应用。

A. 统计报表　B. 网络爬虫　C. API 接口　D. 传感器

9. 下列关于数据重组的说法中,错误的是(　　)。

A. 数据重组是数据的重新生产和重新采集

B. 数据重组能够使数据焕发新的光芒

C. 数据重组实现的关键在于多源数据融合和数据集成

D. 数据重组有利于实现新颖的数据模式创新

10. 下列关于脏数据的说法中,正确的是(　　)。(多选题)

A. 格式不规范　B. 编码不统一

C. 意义不明确　D. 与实际业务关系不大

E. 数据不完整

11. 采样分析的精确性随着采样随机性的增加而(　　),但与样本数量的增加关系不大。

A. 降低　B. 不变　C. 提高　D. 无关

12. 将原始数据进行集成、变换、维度规约、数值规约是在以下哪个步骤的任务?(　　)

A. 频繁模式挖掘　B. 分类和预测

C. 数据预处理　D. 数据流挖掘

二、问答题

1. 简述大数据采集的概念。

2. 绘出数据采集工作流程图。
3. 简述大数据导入/预处理的过程。
4. 什么是数据清洗?
5. 简述数据采集(ETL)技术。
6. 分别描述异构数据交换方式和技术。

第4章　大数据存储

Web、移动设备和其他技术的出现导致数据性质的根本性变化。大数据具有重要而独特的特性,这种特性使得它与“传统”企业数据区分开来。不再集中化、高度结构化并且易于管理,与以往任何时候相比,现在的数据都是高度分散的、结构松散(如果存在结构的话)并且体积越来越大。传统数据与大数据的特性比较见表4.1。

表4.1　传统数据与大数据对比

传统数据	大数据
千兆字节～百万兆字节	拍字节(PB)～艾字节(EB)
集中化	分布式
结构化	半结构化和无结构化
稳定的数据模型	平面模型
已知的复杂的内部关系	不复杂的内部关系

从时间或成本效益上看,传统的数据仓库等数据管理工具都无法实现大数据的处理和分析工作。也就是说,必须将数据组织成关系表(整齐的行和列数据),传统的企业级数据仓库才可以处理。由于需要的时间和人力成本,对海量的非结构化数据应用这种结构是不切实际的。此外,要扩展传统的企业级数据仓库使其适应潜在的PB级数据,需要在新的专用硬件上投资巨额资金。而由于数据加载这一瓶颈,传统数据仓库性能也会受到影响。

因此,需要存储大数据的新方法。

4.1　传统数据存储

4.1.1　传统数据存储介质

数据存储介质分为磁带、磁盘和光盘三大类,由三种介质分别构成的磁带库、磁盘阵列、光盘库三种主要存储设备,三种不同的存储介质具有不同的数据存储特点(见表4.2)。

目前市场上的存储产品主要有磁盘阵列、磁带机与磁带库、光盘库等,其中磁盘设备由于存取速度快、数据查询方便、简单易用、安全的RAID技术等占据一级存储市场的主要份额,磁带设备则以技术成熟、价格低廉等优点占据了二级存储市场的重要地位,光盘设备由于同时具有二者的特点,因此应用在广泛的领域中。

表 4.2　存储介质种类及特点

介质分类	介 质 优 点	介质缺点	数据存储速度	应用环境
磁带	容量大、保存时间长	数据顺序检索，定位时间长	慢	海量数据的定期备份
磁盘	数据读取、写入速度快，操作方便	发热量大、噪声大、硬盘易损	很快	海量数据的即时存取
光盘	单位存储容量成本低，携带方便，数据查询时间短	表面易磨损、寿命短	快	海量数据的在线访问和离线存储

1. 磁带库存储

自从第一台磁带驱动器 IBM726 发明以后，磁带存储技术经过了多年的发展，具有稳定、高可用、低成本等诸多优点，磁带已经成为重要的存储设备。磁带技术可以通过脱机来避免在数据备份、迁移和保护等应用中数据丢失的可能性。另外，磁带技术在高可靠性、低成本等方面也比其他存储设备具有优势，至今相同容量的磁带库成本比磁盘的 RAID 系统还是要低很多，因此只要不断提高 I/O 的传输速率，增加单个磁带的数据容量，简化磁带管理软件的应用界面，磁带技术就不会在短期内过时，目前解决企业数据长期保存的有效方法依然是采用磁带存储技术。随着制造技术和生产工艺的不断改进，磁带将被做得越来越小，存储能力越来越大，磁带库所占空间将减小。随着磁带机的自动化程度的提高，传动系统故障率的降低，磁带存储性能的提高，磁带在存储备份市场的主导地位还会保持相当长的时间。

2. 光盘海量存储

光盘存储技术是近年来发展迅速的光学信息存储新技术。光盘存储技术是一种光学信息存储技术，通过调制激光束在光学圆盘镀膜介质中把信息编码以光点的形式记录下来。在记录及读取过程中，激光头不直接接触光盘的表面，光盘上的记录信息不易被破坏，具有存储密度高、容量大、检索时间短、易于复制、保存时间长、应用领域广等诸多优点，因此光盘海量存储技术被大量的应用。

单张光盘的存储容量从 CD 盘片的几百兆字节到最新的蓝光 DVD 几十吉字节，这样的容量对于海量信息存储系统来讲是远远不够的，要想获得海量的数据存取，就必须将大量存储不同信息的几十、上百甚至上千张光盘组合起来使用。光盘存储的主要形式有以下几种：光盘塔、SCSI 光盘塔、网络光盘塔、光盘库、光盘镜像服务器（见表 4.3），其中光盘网络镜像服务器是一种网络附加存储设备，代表了光盘库的发展方向。

表 4.3　三种光盘设备性能比较表

设备分类	访问速度	容量	成本	可共享用户数	应用环境
光盘塔	中等	小	较高	少	片库
光盘库	慢	较大	最高	少	图书馆、信息管理中心
光盘镜像服务器	很快	最大	最低	多	多种网络环境

随着光存储技术的发展，光盘产品不断的系列化，光存储设备价格不断的降低，应用领域越来越广泛，不仅满足海量数据的存储还能实现一些基本的离线备份功能，因此目前多媒体海量信息存储载体或重要文献资料备份媒体仍然采用光盘介质。

当然光盘技术也存在着一些不足之处，还有一些尚待研究和解决的问题，例如记录速度慢、保存时间短等，另外光盘存储格式还未建立统一的光盘技术国际标准。随着记录介质、记录方法和系统性能的不断改进和提高，光盘存储技术一定会达到更加完善的程度，从而不断满足人们对海量信息存储新的要求。

3. 磁盘阵列海量存储

磁盘阵列又称为廉价磁盘冗余阵列(Redundant Array of Inexpensive Disks，RAID)，是指使用两个或两个以上同类型、容量、接口的磁盘，在磁盘控制器的管理下按照特定的方式组成特定的磁盘组合，从而能快速、准确和安全地读写磁盘数据。

磁盘阵列的特点是将数据有选择性地分布在多个磁盘上，不仅提高数据的可用性及存储容量，而且使得数据存取速度快、吞吐量大，从而避免硬盘故障所带来的灾难后果。磁盘阵列把多个硬盘驱动器连接在一起协同工作，提高了存取速度，同时把磁盘系统的可靠性提高到接近于无错的等级，因此磁盘阵列是一种安全性高，速度快，容量大的存储设备。针对不同的应用磁盘阵列具有多种不同级别，详见表 4.4。

表 4.4 常用 RAID 级别特性比较

RAID 级别	名 称	速 度	容错	磁盘数量	应 用
Level 0	无容错条带磁盘阵列	磁盘并行输入输出	无	至少两块	视频、图像编辑及需要高带宽的应用
Level 1	磁盘镜像方式	读取速度是单个磁盘两倍，写入速度与单个磁盘相同	有	至少两块	会计、金融、付款等需要高可靠性的应用
Level 5	交叉存取加分布奇偶校检	最快的读取速度，中等的写入速度	有	至少三块	文件、数据库 Web、E-mail 等应用服务器
Level 10	镜像条带集	同 Level 0	有	至少四块	数据库服务器和需要高可靠、高性能服务器
Level 0+1	条带集镜像	同 Level 1	有	至少四块	图形应用、通用文件服务器

4.1.2 存储的模式

数据存储需要系统具有良好的数据容错性能和系统稳定性，在发生部分数据错误时，系统可以在线恢复和重建数据，而不影响系统的正常运行。

1. 直连式存储

直连式存储(DAS)即磁盘驱动器和服务器直接连接，存储作为外围设备，在这种存储结构中，数据管理是以服务器为中心的，而且所有的应用软件都是和存储子系统配套的。DAS 适用于一个或有限的几个服务器环境，但存储容量增加时，不但存储供应的效率变得越来越低，而且可升级和扩展性受到很大限制，当服务器出现异常时，更使数据不

可获得,同时存储资源和数据也无法进行共享。

2. 网络存储

网络存储分为网络附加存储(Network Attached Storage,NAS)、光纤存储区域网 FC-SAN、IP 存储区域网 IP-SAN。

NAS 将存储设备连接到现有的网络上来提供数据和文件服务。NAS 服务器一般由存储硬件、操作系统以及其上的文件系统等几个部分组成。NAS 通过网络直接连接磁盘阵列,磁盘阵列具备了高容量、高效能、高可靠等特征。NAS 将存储设备通过标准的网络拓扑结构连接,可以无须服务器直接上网,不依赖通用的操作系统,而是采用一个面向用户设计的、专门用于数据存储的简化操作系统,内置与网络连接所需的协议,从而使整个系统的管理和设置较为简单。

光纤存储区域网 FC-SAN 指的是通过一个单独的高速光纤网络把存储设备和挂在 TCP/IP 网络上的服务器群相连。当有海量数据的存取需求时,数据可以通过存储区域网在相关服务器和后台存储设备之间高速传输。SAN 以光纤通道为基础,不但提供了主机和存储设备之间的高速互联,实现了存储设备的共享,服务器通过存储网络直接同存储设备交换数据,不占用 LAN 的网络资源。

IP-SAN 由于主要部分采用光纤通道,设备高昂的成本问题一直未能得到解决,为此将 iSCSI 卡集成到 NAS 存储设备上,支持数据块形式的 I/O 访问,最后发展成主机通过带 TCP 卸载引擎(TCP Off-load Engine,TOE)的 iSCSI 主机总线适配器(Host Bus Adapter,HBA)卡接入 IP 网络来访问 iSCSI 存储设备。IP 存储采用基于 IP 协议的网络传输数据,由于 IP 环境下数据包可以被捕捉解码,对此 iSCSI 存储要采用多种安全措施以提高数据访问和数据存储的安全性。

3. 数据虚拟存储

虚拟存储是将各种存储物理设备整合为一个整体,从而实现在公共控制平台下集中存储资源,统一存储设备的管理,方便用户的数据操作,简化复杂的存储管理配置,使系统能够提供完整、便捷的数据存储功能。虚拟存储技术在用户操作系统看到的存储设备与实际物理存储设备之间搭建了一个虚拟的操作平台,这样从应用程序一直到最终的数据端都可以实施虚拟存储,虚拟化技术的最终功能可以在服务器、网络和存储设备这三个层面上实现,即主机、网络和存储设备三个部分都可实施虚拟存储。

采用虚拟存储技术,可以支持物理磁盘空间动态扩展,从而使用户不必抛弃现有设备,并实现了存储容量的动态扩展。虚拟存储使得数据存储总体成本降低,随着用户对数据管理需求的不断增加,虚拟化技术正在逐步成为存储领域的核心,虚拟存储不仅可以降低存储资源管理的复杂性,而且可以带给系统高可用性和高可靠性,从而降低数据存储管理成本。

4.2 海量数据存储的需求

随着信息社会的发展,越来越多的信息被数据化,尤其是伴随着 Internet 的发展,数

据呈爆炸式增长。从存储服务的发展趋势来看,一方面,是对数据的存储量的需求越来越大;另一方面,是对数据的有效管理提出了更高的要求。首先是存储容量的急剧膨胀,从而对于存储服务器提出了更大的需求;其次是数据持续时间的增加;最后,对数据存储的管理提出了更高的要求。数据的多样化、地理上的分散性、对重要数据的保护等等都对数据管理提出了更高的要求。

随着数字图书馆、电子商务、多媒体传输等用的不断发展,数据从 GB、TB 到 PB 量级海量急速增长。存储产品已不再是附属于服务器的辅助设备,而成为互联网中最主要的花费所在。海量存储技术已成为继计算机浪潮和互联网浪潮之后的第三次浪潮,磁盘阵列与网络存储成为先锋。

1. 海量数据存储简介

海量存储的含义在于,数据存储中的容量增长是没有止境的。因此,用户需要不断地扩张存储空间。但是,存储容量的增长往往同存储性能并不成正比。这也就造成了数据存储上的误区和障碍。

海量存储技术的概念已经不仅仅是单台的存储设备。而多个存储设备的连接使得数据管理成为一大难题。因此,统一平台的数据管理产品近年来受到了 广大用户的欢迎。这一类型的产品能够将不同平台的存储设备整合在一个单一的控制界面上,结合虚拟化软件对存储资源进行管理。这样的产品无疑简化了用户的管理。

数据容量的增长是无限的,如果只是一味地添加存储设备,那么无疑会大幅增加存储成本。因此,海量存储对于数据的精简也提出了要求。同时,不同应用对于存储容量的需求也有所不同,而应用所要求的存储空间往往并不能得到充分利用,这也造成了浪费。

针对以上的问题,重复数据删除和自动精简配置两项技术在近年来受到了广泛的关注和追捧。重复数据删除通过文件块级的比对,将重复的数据块删除而只留下单一实例。这一做法使得冗余的存储空间得到释放,从客观上增加了存储容量。

2. 处理海量数据存储中存在的问题

目前大数据存储面临几个问题:一是存储数据的成本在不断地增加,如何削减开支节约成本以保证高可用性;二是数据存储容量爆炸性增长且难以预估;三是越来越复杂的环境使得存储的数据无法管理。企业信息架构如何适应现状去提供一个较为理想的解决方案,目前业界有几个发展方向。

1)存储虚拟化

对于存储面临的难题,业界采用的解决手段之一就是存储虚拟化。虚拟存储的概念实际上在早期的计算机虚拟存储器中就已经很好地得以体现,常说的网络存储虚拟化只不过是在更大规模范围内体现存储虚拟化的思想。该技术通过聚合多个存储设备的空间,灵活部署存储空间的分配,从而实现现有存储空间高利用率,避免了不必要的设备开支。

存储虚拟化的好处显而易见,可实现存储系统的整合,提高存储空间的利用率,简化系统的管理,保护原有投资等。越来越多的厂商正积极投身于存储虚拟化领域,比如数据复制、自动精简配置等技术也用到了虚拟化技术。

虚拟化并不是一个单独的产品，而是存储系统的一项基本功能。它对于整合异构存储环境、降低系统整体拥有成本是十分有效的。在存储系统的各个层面和不同应用领域都广泛使用虚拟化这个概念。考虑整个存储层次大体分为应用、文件和块设备三个层次，相应的虚拟化技术也大致可以按这三个层次分类。

目前大部分设备提供商和服务提供商都在自己的产品中包含存储虚拟化技术，使得用户能够方便地使用。

2）容量扩展

目前，在发展趋势上，存储管理的重点已经从对存储资源的管理转变到对数据资源的管理。随着存储系统规模的不断扩大，数据如何在存储系统中进行时空分布成为保证数据的存取性能、安全性和经济性的重要问题。面对信息海量增长对存储扩容的需求，目前主流厂商均提出了各自的解决方案。

由于存储现状比较复杂，存储技术的发展业界还没有形成统一的认识，因此在应对存储容量增长的问题上，尚存在很大的提升空间。技术是发展的，数据的世界也是在不断变化的过程中走向完美。企业信息架构的“分”与“合”的情况并不绝对。目前，出现了许多的融合技术，如 NAS 与 SAN 的融合、统一存储网等等。这些都将对企业信息架构产生不同的影响。至于到底采用哪种技术更合适，取决于企业自身对数据的需求。

3. 海量数据存储技术

为了支持大规模数据的存储、传输与处理，针对海量数据存储目前主要开展如下三个方向的研究。

1）虚拟存储技术

存储虚拟化的核心工作是物理存储设备到单一逻辑资源池的映射，通过虚拟化技术，为用户和应用程序提供了虚拟磁盘或虚拟卷，并且用户可以根据需求对它进行任意分割、合并、重新组合等操作，并分配给特定的主机或应用程序，为用户隐藏或屏蔽了具体的物理设备的各种物理特性。存储虚拟化可以提高存储利用率，降低成本，简化存储管理，而基于网络的虚拟存储技术已成为一种趋势，它的开放性、扩展性、管理性等方面的优势将在数据大集中、异地容灾等应用中充分体现出来。

2）高性能 I/O

集群由于其很高的性价比和良好的可扩展性，近年来在 HPC 领域得到了广泛的应用。数据共享是集群系统中的一个基本需求。当前经常使用的是网络文件系统 NFS 或者 CIFS。当一个计算任务在 Linux 集群上运行时，计算结点首先通过 NFS 协议从存储系统中获取数据，然后进行计算处理，最后将计算结果写入存储系统。在这个过程中，计算任务的开始和结束阶段数据读写的 I/O 负载非常大，而在计算过程中几乎没有任何负载。当今的 Linux 集群系统处理能力越来越强，动辄达到几十甚至上百个 TFLOPS，于是用于计算处理的时间越来越短。但传统存储技术架构对带宽和 I/O 能力的提高却非常困难且成本高昂。这造成了当原始数据量较大时，I/O 读写所占的整体时间就相当可观，成为 HPC 集群系统的性能瓶颈。I/O 效率的改进，已经成为今天大多数 Linux 并行集群系统提高效率的首要任务。

3）网格存储系统

高能物理的数据需求除了容量特别大之外，还要求广泛的共享。比如运行于 BECPII 上的新一代北京谱仪实验 BESIII，未来五年内将累积数据 5PB，分布在全球 20 多个研究单位将对其进行访问和分析。因此，网格存储系统应该能够满足海量存储、全球分布、快速访问、统一命名的需求。主要研究的内容包括网格文件名字服务、存储资源管理、高性能的广域网数据传输、数据复制、透明的网格文件访问协议等。

4. 海量数据处理问题分析

（1）数据量过大，数据中什么情况都可能存在。处理海量数据时，由于软件与硬上都具有很高的要求，可能会造成系统崩溃和硬件损坏，将导致处理程序终止。

（2）软硬件要求高，系统资源占用率高。对海量的数据进行处理，除了好的方法，最重要的就是合理使用工具，合理分配系统资源。一般情况，如果处理的数据在 TB 级以上，小型机是要考虑的，普通的机器如果有好的方法可以考虑，不过也必须加大 CPU 和内存，就像面对着千军万马，光有勇气没有一兵一卒是很难取胜的。

（3）要求很高的处理方法和技巧。好的处理方法是一位工程师长期工作经验的积累，也是个人的经验的总结。没有通用的处理方法，但有通用的原理和规则。

5. 海量数据存储的处理方法

（1）选用优秀的数据库工具。

（2）编写优良的程序代码。

（3）对海量数据进行分区操作。

（4）建立广泛的索引。

（5）建立缓存机制。

（6）加大虚拟内存。

（7）分批处理。

（8）使用临时表和中间表。

（9）优化查询 SQL 语句。

（10）使用文本格式进行处理。

（11）定制强大的清洗规则和出错处理机制。

（12）建立视图或者物化视图。

（13）避免使用 32 位机（极端情况）。

（14）考虑操作系统问题。

（15）使用数据仓库和多维数据库存储。

（16）使用采样数据，进行数据挖掘。

（17）海量数据关联存储。

6. 海量数据是发展前景

海量数据存储技术的发展前展，可以归结为以下几个方面。

1）高容量光存储技术

高容量光存储技术的到来可以说改变了目前的存储格局，为原本暗淡的光存储带来

了一线生机。虽然光存储器的支持者们一直宣传该技术将成为下一代伟大的存储技术，但是即便在它得到推广之后，其企业客户基础在整个市场上的份额仍然很小。

2）分布式存储与 P2P 存储

分布式存储概念提出较早，目前再次成为热点。P2P 存储可以看作分布式存储的一种，是一个用于对等网络的数据存储系统，它的目标是提供高效率的、鲁棒和负载平衡的文件存取功能。

3）数据网格

为了满足人们对高性能、大容量分布存储能力的要求所提出的概念，类似于计算网格，是有机的智能单元的组合。

4）智能存储系统

智能存储系统包括主动的信息采集、主动信息分析、主动调整等。

5）存储服务质量 QoS

应用环境越来越复杂，存储需求区别也越来越明显，这就需要为应用提供区分服务。目前的研究以基于网络存储的 QoS 为主。

6）存储容灾

通过特定的容灾机制，能够在各种灾难损害发生后，最大限度地保障计算机信息系统不间断提供正常应用服务。

7. 大数据存储生命周期过程

大数据分析相比于传统的数据仓库应用，具有数据量大、查询分析复杂等特点。大数据存储由于其本身存在的 4V 特征，传统的存储技术不能满足大数据存储的需要，通过数据采集（ETL）技术数据资源被从源系统中提取，并被转换为一个标准的格式，再使用 NoSQL 数据库进行数据库存取管理，充分利用网络云存储技术节约企业存储成本、提高效率的优势，通过分布式网络文件系统将数据信息存储在整个互联网络资源中，并用可视化的操作界面随时满足用户的数据处理需求。

大数据技术是一个整体，没有统一的解决方案，从大数据生命周期过程的角度可分为数据采集 ETL 技术、NoSQL、云存储、分布式系统、数据可视化 5 个部分。

4.3　分布式存储系统

4.3.1　分布式存储系统

随着全球非结构化数据快速增长，针对结构化数据设计的这些传统存储结构在性能、可扩展性等方面都难以满足要求，进而出现了集群存储、集群并行存储、P2P 存储、面向对象存储等多种存储结构。

1. 集群存储

集群存储，简言之就是将若干个普通性能的存储系统联合起来组成“存储的集群”。集群存储采用开放式的架构，具有很高扩展性，一般包括存储结点、前端网络、后端网络三

个构成元素，每个元素都可以非常容易地进行扩展和升级，而不用改变集群存储的架构。集群存储通过分布式操作系统的作用，会在前端和后端都实现负载均衡。

2. 集群并行存储

集群并行存储采用了分布式文件系统混合并行文件系统。并行存储容许客户端和存储直接打交道，这样可以极大地提高性能。集群并行存储提高了并行或分区 I/O 的整体性能，特别是读取操作密集型以及大型文件的访问。获取更大的命名空间或可编址的阵列。通过在相互独立的存储设备上复制数据来提高可用性。通过廉价的集群存储系统来大幅降低成本，并解决扩展性方面的难题。集群存储多在大型数据中心或高性能计算中心使用。

3. P2P 存储

用 P2P 的方式在广域网中构建大规模分布式存储系统。从体系结构来看，系统采用无中心结构，结点之间对等，通过互相合作来完成用户任务。用户通过该平台自主寻找其他结点进行数据备份和存储空间交换，为用户构建了大规模存储交换的系统平台。P2P 存储用于构建更大规模的分布式存储系统，可以跨多个大型数据中心或高性能计算中心使用。

4. 面向对象存储

面向对象存储是 SAN 和 NAS 的有机结合，是一种存储系统的发展趋势。在面向对象存储中，文件系统中的用户组件部分基本与传统文件系统相同，而将文件系统中的存储组件部分下移到智能存储设备上，于是用户对于存储设备的访问接口由传统的块接口变为对象接口。

4.3.2 典型系统

基于多种分布式文件系统的研究成果，人们对体系结构的认识不断深入，分布式文件系统在体系结构、系统规模、性能、可扩展性、可用性等方面经历了较大的变化。下面按时间顺序介绍几个分布式文件系统的典型应用。

1. NFS

1985 年出现的 NFS 受到了广泛的关注和认可，被移植到了几乎所有主流的操作系统，成为分布式文件系统事实上的标准。NFS 利用 UNIX 系统中的虚拟文件系统(Virtual File System，VFS)机制，将客户机对文件系统的请求，通过规范的文件访问协议和远程过程调用，转发到服务器端进行处理；服务器端在 VFS 之上，通过本地文件系统完成文件的处理，实现了全局的分布式文件系统。Sun 公司公开了 NFS 的实施规范，互联网工程任务组(The Internet Engineering Task Force，IETF)将其列为征求意见稿(Request for Comments，RFC)，这在很大程度上促使 NFS 的很多设计实现方法成为标准，也促进了 NFS 的流行。

2. GPFS

General Parallel File System(GPFS)是目前应用范围较广的一个系统，在系统设计

中采用了多项当时较为先进的技术。GPFS 的磁盘数据结构可以支持大容量的文件系统和大文件，通过采用分片存储、较大的文件系统块、数据预读等方法获得了较高的数据吞吐率；采用扩展哈希(extensible hashing)技术来支持含有大量文件和子目录的大目录，提高文件的查找和检索效率。

GPFS 采用不同粒度的分布式锁解决系统中的并发访问和数据同步问题：字节范围的锁用于用户数据的同步，动态选择元数据结点(metanode)进行元数据的集中管理；具有集中式线索的分布式锁管理整个系统中空间分配等。GPFS 采用日志技术对系统进行在线灾难恢复。每个结点都有各自独立的日志，且单个结点失效时，系统中的其他结点可以代替失效结点检查文件系统日志，进行元数据恢复操作。

GPFS 还有效地克服了系统中任意单个结点的失效、网络通信故障、磁盘失效等异常事件。此外，GPFS 支持在线动态添加、减少存储设备，然后在线重新平衡系统中的数据。这些特性在需要连续作业的高端应用中尤为重要。

3. Storage Tank

IBM 公司在 GPFS 的基础之上发展进化来的 Storage Tank 以及基于 Storage Tank 的 TotalStorage SAN File System 又将分布式文件系统的设计理念和系统架构向前推进了一步。它们除了具有一般的分布式文件系统的特性之外，采用 SAN 作为整个文件系统的数据存储和传输路径。它们采用带外(out-of-band)结构，将文件系统元数据在高速以太网上传输，由专门的元数据服务器来处理和存储。文件系统元数据和文件数据的分离管理和存储，可以更好地利用各自存储设备和传输网络的特性，提高系统的性能，有效降低系统的成本。

Storage Tank 采用积极的缓存策略，尽量在客户端缓存文件元数据和数据。即使打开的文件被关闭，都可以在下次使用时利用已经缓存的文件信息，整个文件系统由管理员按照目录结构划分成多个文件集(fileset)。每一个文件集都是一个相对独立的整体，可以进行独立的元数据处理和文件系统备份等。不同的文件集可以分配到不同的元数据服务器处理，形成元数据服务器机群，提供系统的扩展性、性能、可用性等。

在 TotalStorage 中，块虚拟层将整个 SAN 的存储进行统一的虚拟管理，为文件系统提供统一的存储空间。这样的分层结构有利于简化文件系统的设计和实现。同时，它们的客户端支持多种操作系统，是一个支持异构环境的分布式文件系统。在 SAN File System，采用了基于策略的文件数据位置选择方法，能有效地利用系统的资源、提高性能、降低成本。

4. GFS

GFS(Google File System)系统集群由一个 master 结点和大量的 chunkserver 结点构成，并被许多客户(Client)访问。GFS 把文件分成 64MB 的块，减少了元数据的大小，使 Master 结点能够非常方便地将元数据放置在内存中以提升访问效率。数据块分布在集群的机器上，使用 Linux 的文件系统存放，同时每块文件至少有 3 份以上的冗余。考虑到文件很少被删减或者覆盖，文件操作以添加为主，充分考虑了硬盘线性吞吐量大和随机读点慢的特点。

中心是一个 Master 结点，根据文件索引，找寻文件块。系统保证每个 Master 都会有相应的复制品，以便于在 Master 结点出现问题时进行切换。在 Chunk 层，GFS 将结点失败视为常态，能够非常好地处理 Chunk 结点失效的问题。对于那些稍旧的文件，可以通过对它进行压缩，来节省硬盘空间，且压缩率惊人，有时甚至可以接近 90%。为了保证大规模数据的高速并行处理，引入了 MapReduce 编程模型，同时，由于 MapReduce 将很多烦琐的细节隐藏起来，也极大地简化了程序员的开发工作。

5. Hadoop

Yahoo 也推出了基于 MapReduce 的开源版本 Hadoop，目前 Hadoop 在业界已经被大规模使用。HDFS(Hadoop Distributed File System)有着高容错性的特点，并且设计用来部署在低廉的硬件上，实现了异构软硬件平台间的可移植性。为了尽量减小全局的带宽消耗读延迟，HDFS 尝试返回给一个读操作离它最近的副本。硬件故障是常态，而不是异常，自动地维护数据的多份复制，并且在任务失败后能自动地重新部署计算任务，实现了故障的检测和自动快速恢复。HDFS 放宽了可移植操作系统接口(Portable Operating System Interface，POSIX)的要求，这样可以流的形式访问文件系统中的数据，实现了以流的形式访问写入的大型文件的目的，重点是在数据吞吐量，而不是数据访问的反应时间。

HDFS 提供了接口，来让程序将自己移动到离数据存储更近的位置，消除了网络的拥堵，提高了系统的整体吞吐量。HDFS 的命名空间是由名字结点来存储的。名字结点使用叫做 EditLog 的事务日志来持久记录每一个对文件系统元数据的改变。名字结点在本地文件系统中用一个文件来存储这个 EditLog。整个文件系统命名空间，包括文件块的映射表和文件系统的配置都存在一个叫 FsImage 的文件中，FsImage 也存放在名字结点的本地文件系统中。FsImage 和 Editlog 是 HDFS 的核心数据结构。

4.4 云存储

面对大数据的海量异构数据，传统存储技术面临建设成本高、运维复杂、扩展性有限等问题，成本低廉、提供高可扩展性的云存储技术日益得到关注。

1. 定义

由于业内没有统一的标准，各厂商的技术发展路线也不尽相同，因此相对于云计算，云存储概念存在更多的多义和模糊现象结合云存储技术发展背景及主流厂商的技术方向，可以得出如下定义：云存储是通过集群应用、网格技术或分布式文件系统等，将网络中大量各种不同的存储设备通过应用软件集合起来协同工作，共同对外提供数据存储和业务访问功能的一个系统。

2. 云存储架构

云存储是由一个网络设备、存储设备、服务器、应用软件、公用访问接口、接入网和客户端程序等组成的复杂系统。以存储设备为核心，通过应用软件来对外提供数据存储和业务访问服务。云存储的架构如图 4.1 所示。

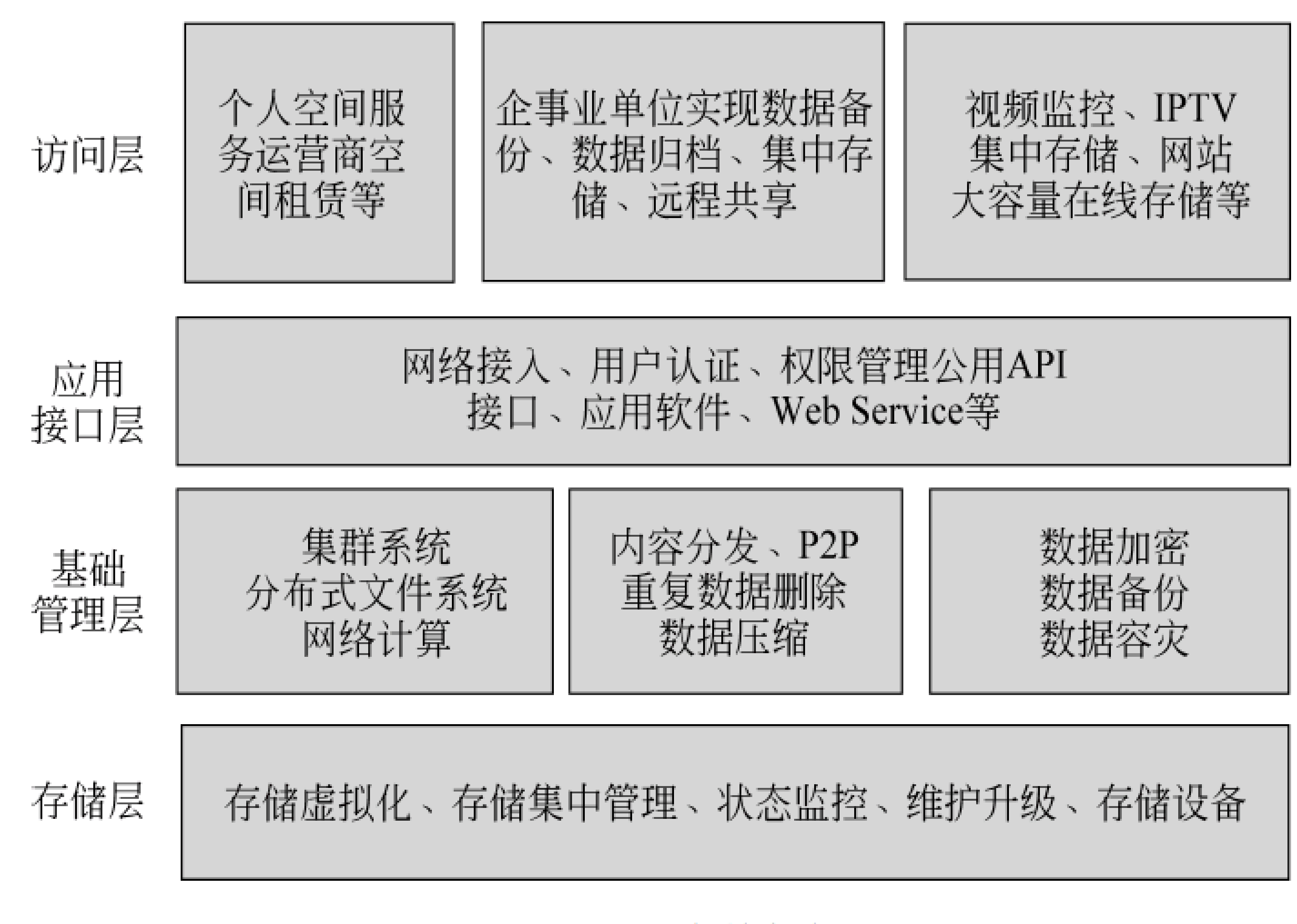


图 4.1　云存储架构

1）存储层

存储设备数量庞大且分布在不同地域，彼此通过广域网、互联网或光纤通道网络连接在一起。在存储设备之上是一个统一存储设备管理系统，实现存储设备的逻辑虚拟化管理、多链路冗余管理，以及硬件设备的状态监控和故障维护。

2）基础管理层

通过集群、分布式文件系统和网格计算等技术，实现云存储设备之间的协同工作，使多个存储设备可以对外提供同一种服务，并提供更大、更强、更好的数据访问性能。数据加密技术保证云存储中的数据不会被未授权的用户访问，数据备份和容灾技术可以保证云存储中的数据不会丢失，保证云存储自身的安全和稳定。

3）应用接口层

不同的云存储运营商根据业务类型，开发不同的服务接口，提供不同的服务。例如视频监控、视频点播应用平台、网络硬盘，远程数据备份应用等。

4）访问层

授权用户可以通过标准的公用应用接口来登录云存储系统，享受云存储服务。

3. 云存储中的数据缩减技术

大数据时代云存储技术的关键技术主要有云存储中的存储虚拟化、分布式存储技术、数据备份、数据缩减技术、内容分发网络技术、数据迁移、数据容错技术等技术，而其中云存储的数据缩减技术，能够满足海量信息爆炸式增长趋势，在一定程度上节约企业存储成本，提高效率，从而成为人们关注的重点。

1）自动精简配置

传统配置技术为了避免重新配置可能造成的业务中断，常常会过度配置容量。在这种情况下，一旦存储分配给某个应用，就不可能重新分配给另一个应用，由此造成已分配的容量没有得到充分利用，造成资源极大浪费。自动精简配置技术利用虚拟化方法减少物理存储空间的分配，最大限度地提升存储空间利用率，其核心原理是“欺骗”操作系统，

让操作系统认为存储设备中有很大的存储空间,而实际的物理存储空间则没有那么大。自动精简配置技术的应用会减少已分配但未使用的存储容量的浪费,在分配存储空间时,需要多少存储空间系统则按需分配。随着数据存储的信息量越来越多,实际存储空间也可以及时扩展,无须用户手动处理。

2)自动存储分层

自动存储分层技术是存储上减少数据的另外一种机制,主要用来帮助数据中心最大限度地降低成本和复杂性。在过去,进行数据移动主要依靠手工操作,由管理员来判断这个卷的数据访问压力或大或小,迁移的时候也只能一个整卷一起迁移。自动存储分层技术的特点则是其分层的自动化和智能化。利用自动存储分层技术一个磁盘阵列能够把活动数据保留在快速、昂贵的存储上,把不活跃的数据迁移到廉价的低速层上,使用户数据保留在合适的存储层级,减少了存储需求的总量,降低了成本,提升了性能。随着固态存储在当前磁盘阵列中的采用以及云存储的来临,自动存储分层已经成为大数据时代补充内部部署的存储的主要方式。

3)重复数据删除

物理存储设备在使用一段时间后必然会出现大量重复的数据。“重复删除”技术(De-duplication)作为一种数据缩减技术可对存储容量进行优化。它通过删除数据集中重复的数据,只保留其中一份,从而消除冗余数据。使用De-dupe技术可以将数据缩减到原来的1/20～1/50。由于大幅度减少了对物理存储空间的信息量,从而达到减少传输过程中的网络带宽、节约设备成本、降低能耗的目的。重复数据删除技术原理De-dupe按照消重的粒度可以分为文件级和数据块级。可以同时使用两种以上的Hash算法计算数据指纹,以获得非常小的数据碰撞发生概率。具有相同指纹的数据块即可认为是相同的数据块,存储系统中仅需要保留一份。这样,一个物理文件在存储系统中就只对应一个逻辑表示。

4)数据压缩

数据压缩技术是提高数据存储效率最古老最有效的方法之一,可以显著降低待处理和存储的数据量,一般情况下可实现2∶1～3∶1的压缩比,对于随机数据效果更好。其原理就是将收到的数据通过存储算法存储到更小的空间中去。在线压缩(RACE)是最新研发的数据压缩技术,与传统压缩技术不同。对RACE技术来说,不仅能在数据首次写入时进行压缩,以帮助系统控制大量数据在主存中杂乱无章地存储的情形。还可以在数据写入到存储系统前压缩数据,进一步提高存储系统中的磁盘和缓存的性能和效率。

数据压缩中使用的LZS算法基于LZ77实现,主要由两部分构成:滑窗(Sliding Window)和自适应编码(Adaptive Coding),如图4.2所示。压缩处理时,在滑窗中查找与待处理数据相同的块,并用该块在滑窗中的偏移值及块长度替代待处理数据,从而实现压缩编码。如果滑窗中没有与待处理数据块相同的字段,或偏移值及长度数据超过被替代数据块的长度,则不进行替代处理。LZS算法的实现非常简洁,处理比较简单,能够适应各种高速应用。

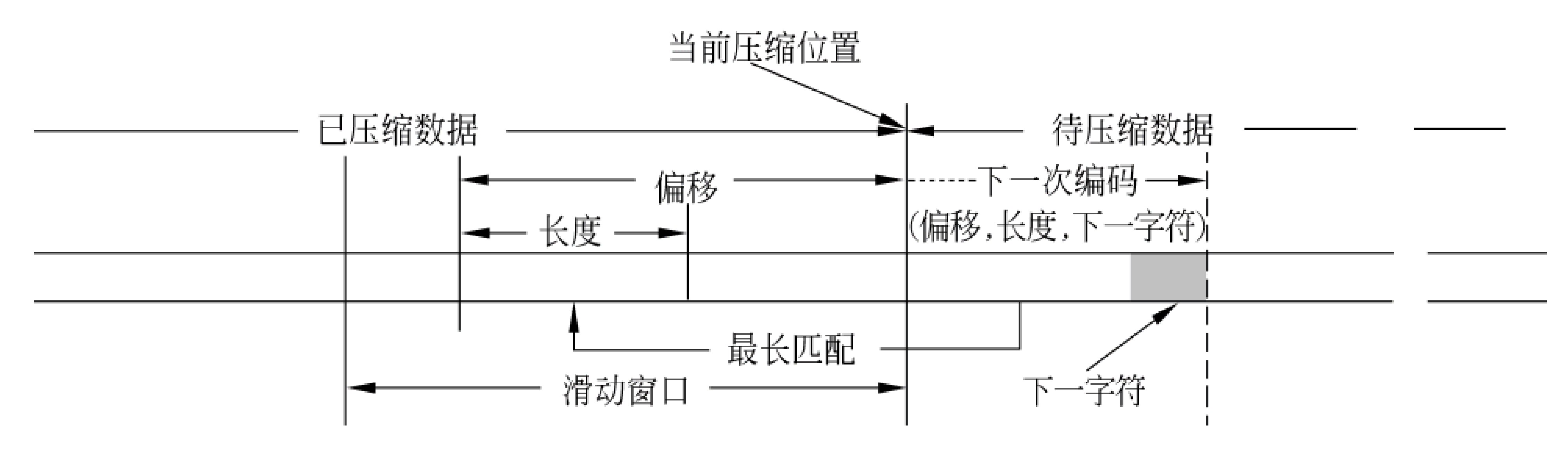


图 4.2　LZ777 算法示意图

4.5　数据库

数据库(Database)是按照数据结构来组织、存储和管理数据的仓库,它产生于距今六十多年前,随着信息技术和市场的发展,特别是 20 世纪 90 年代以后,数据管理不再仅仅是存储和管理数据,而转变成用户所需要的各种数据管理的方式。数据库有很多种类型,从最简单的存储有各种数据的表格到能够进行海量数据存储的大型数据库系统,都在各个方面得到了广泛的应用。

在信息化社会,充分有效地管理和利用各类信息资源,是进行科学研究和决策管理的前提条件。数据库技术是管理信息系统、办公自动化系统、决策支持系统等各类信息系统的核心部分,是进行科学研究和决策管理的重要技术手段。

4.5.1　数据库分类

数据库通常分为层次式数据库、网络式数据库和关系式数据库三种。而不同的数据库是按不同的数据结构来联系和组织的。

1. 数据结构模型

1) 数据结构

所谓数据结构,是指数据的组织形式或数据之间的联系。

如果用 D 表示数据,用 R 表示数据对象之间存在的关系集合,则将 DS=(D,R)称为数据结构。

例如,设有一个电话号码簿,它记录了 n 个人的名字和相应的电话号码。为了方便地查找某人的电话号码,将人名和号码按字典顺序排列,并在名字的后面跟随着对应的电话号码。这样,若要查找某人的电话号码(假定他的名字的第一个字母是 Y),那么只需查找以 Y 开头的那些名字就可以了。该例中,数据的集合 D 就是人名和电话号码,它们之间的联系 R 就是按字典顺序的排列,其相应的数据结构就是 DS=(D,R),即一个数组。

2) 数据结构类型

数据结构又分为数据的逻辑结构和数据的物理结构。

数据的逻辑结构是从逻辑的角度(即数据间的联系和组织方式)来观察数据、分析数据,与数据的存储位置无关;数据的物理结构是指数据在计算机中存放的结构,即数据的

逻辑结构在计算机中的实现形式，所以物理结构也被称为存储结构。

这里只研究数据的逻辑结构，并将反映和实现数据联系的方法称为数据模型。

比较流行的数据模型有三种，即按图论理论建立的层次结构模型、网状结构模型以及按关系理论建立的关系结构模型。

2. 层次、网状和关系数据库系统

1）层次结构模型

层次结构模型实质上是一种有根结点的定向有序树（在数学中“树”被定义为一个无回的连通图）。例如一个高等学校的组织结构就像一棵树，校部就是树根（称为根结点），各系、专业、教师、学生等为枝点（称为结点），树根与枝点之间的联系称为边，树根与边之比为 1∶N，即树根只有一个，树枝有 N 个。

按照层次模型建立的数据库系统称为层次模型数据库系统。IMS（Information Management System）是其典型代表。

2）网状结构模型

按照网状数据结构建立的数据库系统称为网状数据库系统，其典型代表是 DBTG（Database Task Group）。用数学方法可将网状数据结构转化为层次数据结构。

3）关系结构模型

关系式数据结构把一些复杂的数据结构归结为简单的二元关系（即二维表格形式）。例如某单位的职工关系就是一个二元关系。

由关系数据结构组成的数据库系统被称为关系数据库系统。

在关系数据库中，对数据的操作几乎全部建立在一个或多个关系表格上，通过对这些关系表格的分类、合并、连接或选取等运算来实现数据的管理。

因此，可以概括地说，一个关系称为一个数据库，若干个数据库可以构成一个数据库系统。数据库系统可以派生出各种不同类型的辅助文件和建立它的应用系统。

4.5.2 常规 SQL 结构化关系数据库

结构化查询语言（Structured Query Language）简称 SQL（发音：/es kjuːel/），是一种特殊目的的编程语言，是一种数据库查询和程序设计语言，用于存取数据以及查询、更新和管理关系数据库系统；同时也是数据库脚本文件的扩展名。

结构化查询语言是高级的非过程化编程语言，允许用户在高层数据结构上工作。它不要求用户指定对数据的存放方法，也不需要用户了解具体的数据存放方式，所以具有完全不同底层结构的不同数据库系统，可以使用相同的结构化查询语言作为数据输入与管理的接口。结构化查询语言语句可以嵌套，这使它具有极大的灵活性和强大的功能。

结构化查询语言中的五种数据类型：字符型、文本型、数值型、逻辑型和日期型。

4.5.3 NoSQL 非结构化数据库

NoSQL，泛指非关系型的数据库。随着互联网 Web 2.0 网站的兴起，传统的关系数据库在应付 Web 2.0 网站，特别是超大规模和高并发的 SNS 类型的 Web 2.0 纯动态网

站已经显得力不从心，暴露了很多难以克服的问题，而非关系型的数据库则由于其本身的特点得到了非常迅速的发展。NoSQL 数据库的产生就是为了解决大规模数据集合多重数据种类带来的挑战，尤其是大数据应用难题。

NoSQL(Not Only SQL)，意即“不仅仅是 SQL”，是一项全新的数据库革命性运动，早期就有人提出，发展至 2009 年其趋势越发高涨。NoSQL 的拥护者们提倡运用非关系型的数据存储，相对于铺天盖地的关系型数据库运用，这一概念无疑是一种全新的思维的注入。

这种称为 NoSQL 的新形式的数据库(Not Only SQL)像 Hadoop 一样，可以处理大量的多结构化数据。但是，如果说 Hadoop 擅长支持大规模、批量式的历史分析，在大多数情况下(虽然也有一些例外)，NoSQL 数据库的目的是为最终用户和自动化的大数据应用程序提供大量存储在多结构化数据中的离散数据。这种能力是关系型数据库欠缺的，它根本无法在大数据规模维持基本的性能水平。

在某些情况下，NoSQL 和 Hadoop 协同工作。例如，HBase 是流行的 NoSQL 数据库，它仿照 Google 的 BigTable，通常部署在 HDFS(Hadoop 分布式文件系统)之上，为 Hadoop 提供低延迟的快速查找功能。

目前可用的 NoSQL 数据库包括：HBase、Cassandra、MarkLogic、Aerospike、MongoDB、Accumulo、Riak、CouchDB、DynamoDB。

目前大多数 NoSQL 数据库的缺点是：为了性能和可扩展性，它们遵从 ACID(原子性、一致性、隔离性、持久性)原则。许多 NoSQL 数据库还缺乏成熟的管理和监控工具。

1. NoSQL 数据库的四大分类

1) 键值(Key-Value)存储数据库

这一类数据库主要会使用到一个哈希表，这个表中有一个特定的键和一个指针指向特定的数据。Key-Value 模型对于 IT 系统来说的优势在于简单、易部署。但是如果 DBA 只对部分值进行查询或更新的时候，Key-Value 就显得效率低下了。如 Tokyo Cabinet/Tyrant、Redis、Voldemort、Oracle BDB。

2) 列存储数据库

这部分数据库通常是用来应对分布式存储的海量数据。键仍然存在，但是它们的特点是指向了多个列。这些列是由列家族来安排的。如 Cassandra、HBase、Riak。

3) 文档型数据库

文档型数据库的灵感是来自于 Lotus Notes 办公软件的，而且它同第一种键值存储相类似。该类型的数据模型是版本化的文档，半结构化的文档以特定的格式存储，比如 JSON。文档型数据库可以看作是键值数据库的升级版，允许之间嵌套键值。而且文档型数据库比键值数据库的查询效率更高。如 CouchDB、MongoDb。国内也有文档型数据库 SequoiaDB，已经开源。

4) 图形(Graph)数据库

图形结构的数据库同其他行列以及刚性结构的 SQL 数据库不同，它是使用灵活的图形模型，并且能够扩展到多个服务器上。NoSQL 数据库没有标准的查询语言(SQL)，因

此进行数据库查询需要制定数据模型。许多 NoSQL 数据库都有 REST 式的数据接口或者查询 API。如 Neo4J、InfoGrid、Infinite Graph。

2. NoSQL 数据库的四大分类表格分析

常见的 NoSQL 数据库可分为四大类，为便于说明，列出表 4.5。

表 4.5 NoSQL 数据库的四大分类

分类	Examples 举例	典型应用场景	数据模型	优 点	缺 点
键值(key-value)	Tokyo Cabinet/Tyrant，Redis，Voldemort，Oracle BDB	内容缓存，主要用于处理大量数据的高访问负载，也用于一些日志系统等等	Key 指向 Value 的键值对，通常用 hash table 来实现	查找速度快	数据无结构化，通常只被当作字符串或者二进制数据
列存储数据库	Cassandra，HBase，Riak	分布式的文件系统	以列簇式存储，将同一列数据存在一起	查找速度快，可扩展性强，更容易进行分布式扩展	功能相对局限
文档型数据库	CouchDB，MongoDb	Web 应用（与 Key-Value 类似，Value 是结构化的，不同的是数据库能够了解 Value 的内容）	Key-Value 对应的键值对，Value 为结构化数据	数据结构要求不严格，表结构可变，不需要像关系型数据库一样需要预先定义表结构	查询性能不高，而且缺乏统一的查询语法
图形(Graph)数据库	Neo4J，InfoGrid，Infinite Graph	社交网络，推荐系统等。专注于构建关系图谱	图结构	利用图结构相关算法。比如最短路径寻址，*N* 度关系查找等	很多时候需要对整个图做计算才能得出需要的信息，而且这种结构不太好做分布式的集群方案

3. 适用场景

NoSQL 数据库在以下的这几种情况下比较适用：

(1) 数据模型比较简单；

(2) 需要灵活性更强的 IT 系统；

(3) 对数据库性能要求较高；

(4) 不需要高度的数据一致性；

(5) 对于给定 key，比较容易映射复杂值的环境。

4.5.4 NoSQL 技术

在大数据时代，Web 2.0 网站要根据用户个性化信息来实时生成动态页面和提供动态信息，所以基本上无法使用动态页面静态化技术，因此数据库并发负载非常高，往往要达到每秒上万次读写请求。关系数据库应付上万次 SQL 查询还勉强顶得住，但是应付上

万次 SQL 写数据请求,硬盘 I/O 就已经无法承受了。

对于大型的 SNS 网站,每天用户产生海量的用户动态,对于关系数据库来说,在庞大的表里面进行 SQL 查询,效率是极其低下乃至不可忍受的。

此外,在基于 Web 的架构当中,数据库是最难进行横向扩展的,当一个应用系统的用户量和访问量与日俱增的时候,你的数据库却没有办法像 Web Server 和 App Server 那样简单地通过添加更多的硬件和服务结点来扩展性能和负载能力。对于很多需要提供 24 小时不间断服务的网站来说,对数据库系统进行升级和扩展是非常痛苦的事情,往往需要停机维护和数据迁移,为什么数据库不能通过不断地添加服务器结点来实现扩展呢?

所以上面提到的这些问题和挑战都在催生一种新型数据库技术的诞生,这就是 NoSQL 技术。

1. NoSQL 与关系型数据库设计理念比较

关系型数据库中的表都是存储一些格式化的数据结构,每个元组字段的组成都一样,即使不是每个元组都需要所有的字段,但数据库会为每个元组分配所有的字段,这样的结构便于表与表之间进行连接等操作,但从另一个角度来说它也是关系型数据库性能瓶颈的一个因素。而非关系型数据库以键值对存储,它的结构不固定,每一个元组可以有不一样的字段,每个元组可以根据需要增加一些自己的键值对,这样就不会局限于固定的结构,可以减少一些时间和空间的开销。

2. NoSQL 技术特点

1）易扩展性

NoSQL 数据库种类繁多,但是一个共同的特点都是去掉关系数据库的关系型特性。数据之间无关系,这样就非常容易扩展。无形之间,在架构的层面上带来了可扩展的能力。

2）大数据量,高性能

NoSQL 数据库都具有非常高的读写性能,尤其在大数据量下,同样表现优秀。这得益于它的无关系性,数据库的结构简单。一般 MySQL 使用 Query Cache,每次表的更新 Cache 就失效,是一种大粒度的 Cache,在针对 Web 2.0 的交互频繁的应用,Cache 性能不高。而 NoSQL 的 Cache 是记录级的,是一种细粒度的 Cache,所以 NoSQL 在这个层面上来说性能就要高很多了。

3）灵活的数据模型

NoSQL 无须事先为要存储的数据建立字段,随时可以存储自定义的数据格式。而在关系数据库里,增删字段是一件非常麻烦的事情。如果是非常大数据量的表,增加字段简直就是一个噩梦。这点在大数据量的 Web 2.0 时代尤其明显。高可用:NoSQL 在不太影响性能的情况,就可以方便地实现高可用的架构。比如 Cassandra、HBase 模型,通过复制模型也能实现高可用。

3. CAP 原理

分布式数据系统的三要素:一致性(Consistency)、可用性(Availability)和分区容忍性(Partition tolerance)。CAP 原理是指,在分布式系统中,这三个要素最多只能同时实

现两点，不可能三者兼顾。对于分布式数据系统，分区容忍性是基本要求。对于大多数Web应用，牺牲一致性而换取高可用性，是目前多数分布式数据库产品的方向。

4. 几种主流 NoSQL 数据库

而互联网庞大的数据量和极高的峰值访问压力使得以增加内存、CPU等结点性能的垂直伸缩方案(Scale-UP)走入死胡同，使用大量廉价的机器组建水平可扩展集群(Scale Out)成为绝大多数互联网公司的必然选择；廉价的机器失效是正常的，大规模的集群，结点之间的网络临时阻断也是常见的，因此在衡量一致性、可用性和分区容忍性时，往往倾向先满足后两者，再用其他方法满足最终的一致性。在衡量CAP时，Bigtable选择了CA，用GFS来弥补P，Dynamo选择了AP，C弱化为最终一致性(通过Quorum或者read-your-write机制)。

1) BigTable

(1) BigTable简介。

Bigtable是一个分布式的结构化数据存储系统，它被设计用来处理海量数据：通常是分布在数千台普通服务器上的PB级的数据。Google的很多项目使用Bigtable存储数据，包括Web索引、Google Earth、Google Finance等。

(2) 数据模型。

Bigtable是一个稀疏的、分布式的、持久化存储的多维度排序Map。Map的索引是行关键字、列关键字以及时间戳；Map中的每个value都是一个未经解析的byte数组。

一个存储Web网页的例子的表的片断如下：

行名：'com. cnn. www'

contents列族：存放的是网页的内容。

anchor列族：存放引用该网页的锚链接文本。

“anchor:cnnsi. com”列表示被cnnsi. com引用。

“anchhor:my. look. ca”列表示被my. look. ca引用。

(3) 技术要点。

基础：GFS、Chubby、SSTable。

- BigTable使用Google的分布式文件系统(GFS)存储日志文件和数据文件。
- Chubby是一个高可用的、序列化的分布式锁服务组件。
- BigTable内部存储数据的文件是Google SSTable格式的。
- 元数据与数据都保存在Google FS中，客户端通过Chubby服务获得表格元数据的位置。

数据维护与访问：master server将每个tablet的管理责任分配给各个tablet server，tablet的分布信息都保存在元数据中，所以客户端无须通过master来访问数据，只需要直接跟tablet server通信。

Log-structured数据组织：写操作不直接修改原有的数据，而只是将一条记录添加到commit log的末尾，读操作需要从log中merge出当前的数据版本。具体实现：SSTable和Memtable(Memtable即内存表：将新数据或常用数据保存在内存表，可以减少磁盘

IO 访问）。

（4）特点。

- 适合大规模海量数据，PB 级数据；
- 分布式、并发数据处理，效率极高；
- 易于扩展，支持动态伸缩，适用于廉价设备；
- 适合于读操作，不适合写操作；
- 不适用于传统关系数据库。

2）Dynamo

（1）Dynamo 简介。

Dynamo 最初是 Amazon 所使用的一个私有的分布式存储系统。

（2）设计要点。

P2P 的架构：区别于 Google FS 的 Single Master 架构，P2P 架构无须一个中心服务器来记录系统的元数据。可以根据应用的需求自由调整 Performance（性能）、Availability（可用性）、Durability（数据持久性）三者的比例。

（3）技术要点。

将所有主键的哈希数值空间组成一个首位相接的环状序列，对于每台机器，随机赋予一个哈希值，不同的机器就会组成一个环状序列中的不同结点，而该机器就负责存储落在一段哈希空间内的数据。数据定位使用一致性哈希；对于一个数据，首先计算其的哈希值，根据其所落在的某个区段，顺时针进行查找，找到第一台机，该机器就负责存储在数据的，对应的存取操作及冗余备份等操作也有其负责，以此来实现数据在不同机器之间的动态分配。

对于一个环状结点比如 M 个结点，比如一份数据需要保持 N 个备份，则该数据落在某个哈希区间内发现的第一个结点负责后续对应的 $N-1$ 个结点的数据备份（注意 $M>=N$），Vector lock，允许数据的多个备份存在多个版本，提高写操作的可用性（用弱一致性来换取高的可用性）分布式存储系统对于某个数据保存多个备份，数据写入要尽量保证备份数据同时获得更新 Dynamo 采取数据最终一致性，在一定的时间窗口中，对数据的更新会传播到所有备份中，但是在时间窗口内，如果有客户读取到旧的数据，通过向量时钟（Vector Clock）。

4.5.5　大规模并行分析数据库

不同于传统的数据仓库，大规模并行分析数据库能够以必需的最少的数据建模，快速获取大量的结构化数据，可以向外扩展以容纳 TB 甚至 PB 级数据。

对最终用户而言最重要的是，大规模并行分析数据库支持近乎实时的复杂 SQL 查询结果，也叫交互式查询功能，而这正是 Hadoop 显著缺失的能力。大规模并行分析数据库在某些情况下支持近实时的大数据应用。大规模并行分析数据库的基本特性包括如下几个方面。

1. 大规模并行分析数据库的基本特性

1）大规模并行处理的能力

就像其名字表明的一样，大规模并行分析数据库采用大规模并行处理同时支持多台

机器上的数据采集、处理和查询。相对传统的数据仓库具有更快的性能,传统数据仓库运行在单一机器上,会受到数据采集这个单一瓶颈点的限制。

2) 无共享架构

无共享架构可确保分析数据库环境中没有单点故障。在这种架构下,每个结点独立于其他结点,所以如果一台机器出现故障,其他机器可以继续运行。对大规模并行处理环境而言,这点尤其重要,数百台计算机并行处理数据,偶尔出现一台或多台机器失败是不可避免的。

3) 列存储结构

大多数大规模并行分析数据库采用列存储结构,而大多数关系型数据库以行结构存储和处理数据。在列存储环境中,由包含必要数据的列决定查询语句的“答案”,而不是由整行的数据决定,从而导致查询结果瞬间可以得出。这也意味着数据不需要像传统的关系数据库那样构造成整齐的表格。

4) 强大的数据压缩功能

它们允许分析数据库收集和存储更大量的数据,而且与传统数据库相比占用更少的硬件资源。例如,具有 10 比 1 的压缩功能的数据库,可以将 10TB 字节的数据压缩到 1TB。数据编码(包括数据压缩以及相关的技术)是有效地扩展到海量数据的关键。

5) 商用硬件

像 Hadoop 集群一样,大多数(肯定不是全部)大规模并行分析数据库运行在戴尔、IBM 等厂商现成的商用硬件上,这使他们能够以具有成本效益的方式向外扩展。

6) 在内存中进行数据处理

有些(肯定不是全部)大规模并行分析数据库使用动态 RAM 或闪存进行实时数据处理。有些(如 SAP HANA 和 Aerospike)完全在内存中运行数据,而其他则采用混合的方式,即用较便宜但低性能的磁盘内存处理“冷”数据,用动态 RAM 或闪存处理“热”数据。

然而,大规模并行分析数据库确实有一些盲点。最值得注意的是,它们并非被设计用来存储、处理和分析大量的半结构化和非结构化数据。

2. 大数据方法的互补

Hadoop、NoSQL 和大规模并行分析数据库不是相互排斥的。相反,这三种方法是互补的,彼此可以而且应该共存于许多企业。Hadoop 擅长处理和分析大量分布式的非结构化数据,以分批的方式进行历史分析。NoSQL 数据库擅长为基于 Web 的大数据应用程序提供近实时地多结构化数据存储和处理。而大规模并行分析数据库最擅长对大容量的主流结构化数据提供接近实时的分析。

例如,Hadoop 完成的历史分析可以移植到分析数据库供进一步分析,或者与传统的企业数据仓库的结构化数据进行集成。从大数据分析得到的见解可以而且应该通过大数据应用实现产品化。企业的目标应该是实现一个灵活的大数据架构,在该架构中,三种技术可以尽可能无缝地共享数据和见解。

很多预建的连接器可以帮助 Hadoop 开发者和管理员实现这种数据集成，同时也有很多厂商(包括 Pivotal Initiative(原 EMC 的 Greenplum)，CETAS 和 Teradata Aster)提供大数据应用。这些大数据应用将 Hadoop、分析数据库和预配置的硬件进行捆绑，可以达到以最小的调整实现快速部署的目的。另外一种情况，Hadapt 提供了一个单一平台，这个平台在相同的集群上同时提供 SQL 和 Hadoop/MapReduce 的处理功能。Cloudera 也在 Impala 和 Hortonworks 项目上通过开源倡议推行这一策略。

但是，为了充分利用大数据，企业必须采取进一步的措施。也就是说，他们必须使用高级分析技术处理数据，并以此得出有意义的见解。数据科学家通过屈指可数的语言或方法(包括 SAS 和 R)执行这项复杂的工作。分析的结果可以通过 Tableau 这样的工具可视化，也可以通过大数据应用程序进行操作，这些大数据应用程序包括自己开发的应用程序和现成的应用程序。其他厂商(包括 Platfora 和 Datameer)正在开发商业智能型的应用程序，这种应用程序允许非核心用户与大数据直接交互。

底层的大数据方法(如 Hadoop、NoSQL 和大规模并行分析数据库)不仅本身是互补的，而且与大部分大型企业现有的数据管理技术互补。并不建议企业为了大数据方法而“淘汰并更换”企业现有的全部的数据仓库、数据集成和其他数据管理技术。

相反，必须像投资组合经理那样思考，重新权衡优先级，为企业走向创新和发展奠定基础，同时采取必要的措施减轻风险因素。用大数据方法替换现有的数据管理技术，只有当它的商业意义和发展计划与现有的数据管理基础设施尽可能无缝地整合时才有意义。最终目标应该是转型为现代数据架构。

4.6 数据仓库

数据仓库，英文名称为 Data Warehouse，可简写为 DW 或 DWH。数据仓库，是为企业所有级别的决策制定过程，提供所有类型数据支持的战略集合。它是单个数据存储，出于分析性报告和决策支持目的而创建。为需要业务智能的企业提供指导业务流程改进、监视时间、成本、质量以及控制。

数据仓库中的数据是在对原有分散的数据库数据抽取、清理的基础上经过系统加工、汇总和整理得到的，必须消除源数据中的不一致性，以保证数据仓库内的信息是关于整个企业的一致的全局信息。

数据仓库的数据主要供企业决策分析之用，所涉及的数据操作主要是数据查询，一旦某个数据进入数据仓库以后，一般情况下将被长期保留，也就是数据仓库中一般有大量的查询操作，但修改和删除操作很少，通常只需要定期的加载、刷新。

4.6.1 数据仓库的概念

数据仓库是决策支持系统和联机分析应用数据源的结构化数据环境。数据仓库研究和解决从数据库中获取信息的问题。数据仓库的特征在于面向主题、集成性、稳定性和时变性。

数据仓库的概念由“数据仓库之父”比尔 · 恩门(Bill Inmon)于 1990 年提出，其主要

功能仍是将组织透过资讯系统之联机事务处理(OLTP)经年累月所累积的大量资料,通过数据仓库理论所特有的资料储存架构,做一有系统性的分析整理,以利各种分析方法,如联机分析处理(OLAP)、数据挖掘(Data Mining)的进行,并进而支持如决策支持系统(DSS)、主管资讯系统(EIS)的创建,帮助决策者能快速有效地从大量资料中分析出有价值的资讯,以利决策拟定及快速回应外在环境变动,帮助建构商业智能(BI)。

数据仓库是在数据库已经大量存在的情况下,为了进一步挖掘数据资源、为了决策需要而产生的,它并不是所谓的"大型数据库"。数据仓库的方案建设的目的,是为前端查询和分析作为基础,由于有较大的冗余,所以需要的存储也较大。为了更好地为前端应用服务,数据仓库往往有如下特点。

1. 效率足够高

数据仓库的分析数据一般分为日、周、月、季、年等,可以看出,日为周期的数据要求的效率最高,要求 24 小时甚至 12 小时内,客户能看到昨天的数据分析。由于有的企业每日的数据量很大,设计不好的数据仓库经常会出问题,延迟 1～3 日才能给出数据,显然不行的。

2. 数据质量

数据仓库所提供的各种信息,肯定要准确的数据,但由于数据仓库流程通常分为多个步骤,包括数据清洗、装载、查询、展现等等,复杂的架构会更多层次,那么由于数据源有脏数据或者代码不严谨,都可以导致数据失真,客户看到错误的信息就可能导致分析出错误的决策,造成损失,而不是产生效益。

3. 扩展性

之所以有的大型数据仓库系统架构设计复杂,是因为考虑到了未来 3～5 年的扩展性,这样的话,未来不用太快花钱去重建数据仓库系统,就能很稳定地运行。主要体现在数据建模的合理性上,数据仓库方案中多出一些中间层,使海量数据流有足够的缓冲,不至于数据量大很多,就运行不起来了。

从上面的介绍中可以看出,数据仓库技术可以将企业多年积累的数据唤醒,不仅为企业管理好这些海量数据,而且挖掘数据潜在的价值,从而成为通信企业运营维护系统的亮点之一。

从广义上说,基于数据仓库的决策支持系统由三个部件组成:数据仓库技术、联机分析处理技术和数据挖掘技术,其中数据仓库技术是系统的核心。

4. 面向主题

操作型数据库的数据组织面向事务处理任务,各个业务系统之间各自分离,而数据仓库中的数据是按照一定的主题域进行组织的。主题是与传统数据库的面向应用相对应的,是一个抽象概念,是在较高层次上将企业信息系统中的数据综合、归类并进行分析利用的抽象。每一个主题对应一个宏观的分析领域。数据仓库排除对于决策无用的数据,提供特定主题的简明视图。

4.6.2　数据仓库技术发展

企业的数据处理大致分为两类：一类是操作型处理，也称为联机事务处理，它是针对具体业务在数据库联机的日常操作，通常对少数记录进行查询、修改；另一类是分析型处理，一般针对某些主题的历史数据进行分析，支持管理决策。

两者具有不同的特征，主要体现在以下几个方面。

1. 处理性能

日常业务涉及频繁、简单的数据存取，因此对操作型处理的性能要求是比较高的，需要数据库能够在很短时间内做出反应。

2. 数据集成

企业的操作型处理通常较为分散，传统数据库面向应用的特性使数据集成困难。

3. 数据更新

操作型处理主要由原子事务组成，数据更新频繁，需要并行控制和恢复机制。

4. 数据时限

操作型处理主要服务于日常的业务操作。

5. 数据综合

操作型处理系统通常只具有简单的统计功能。

数据库已经在信息技术领域有了广泛的应用，我们社会生活的各个部门，几乎都有各种各样的数据库保存着与我们的生活息息相关的各种数据。作为数据库的一个分支，数据仓库概念的提出，相对于数据库从时间上就近得多。美国著名信息工程专家 Willian H. Inmon 在 20 世纪 90 年代初提出了数据仓库概念的一个表述，认为：“一个数据仓库通常是一个面向主题的、集成的、随时间变化的、但信息本身相对稳定的数据集合，它用于对管理决策过程的支持。”

这里的主题，是指用户使用数据仓库进行决策时所关心的重点方面，如收入、客户、销售渠道等；所谓面向主题，是指数据仓库内的信息是按主题进行组织的，而不是像业务支撑系统那样是按照业务功能进行组织的。

4.6.3　数据仓库原理及构成

1. 数据仓库系统的概念

数据仓库系统是一个系统的工程，而不是一件产品，提供用户用于决策支持的当前和历史的数据(这些数据在传统的操作型数据库中很难或不能得到)，并通过联机分析处理(OLAP)、数据挖掘(DM)和快速报表工具等技术对这些数据进行处理，为决策提供需要的信息。数据仓库技术是为了有效地把操作形数据集成到统一的环境中以提供决策型数据访问，并进行分析、挖掘的各种技术和模块的总称。

图 4.3 描述了一个典型的数据仓库系统。

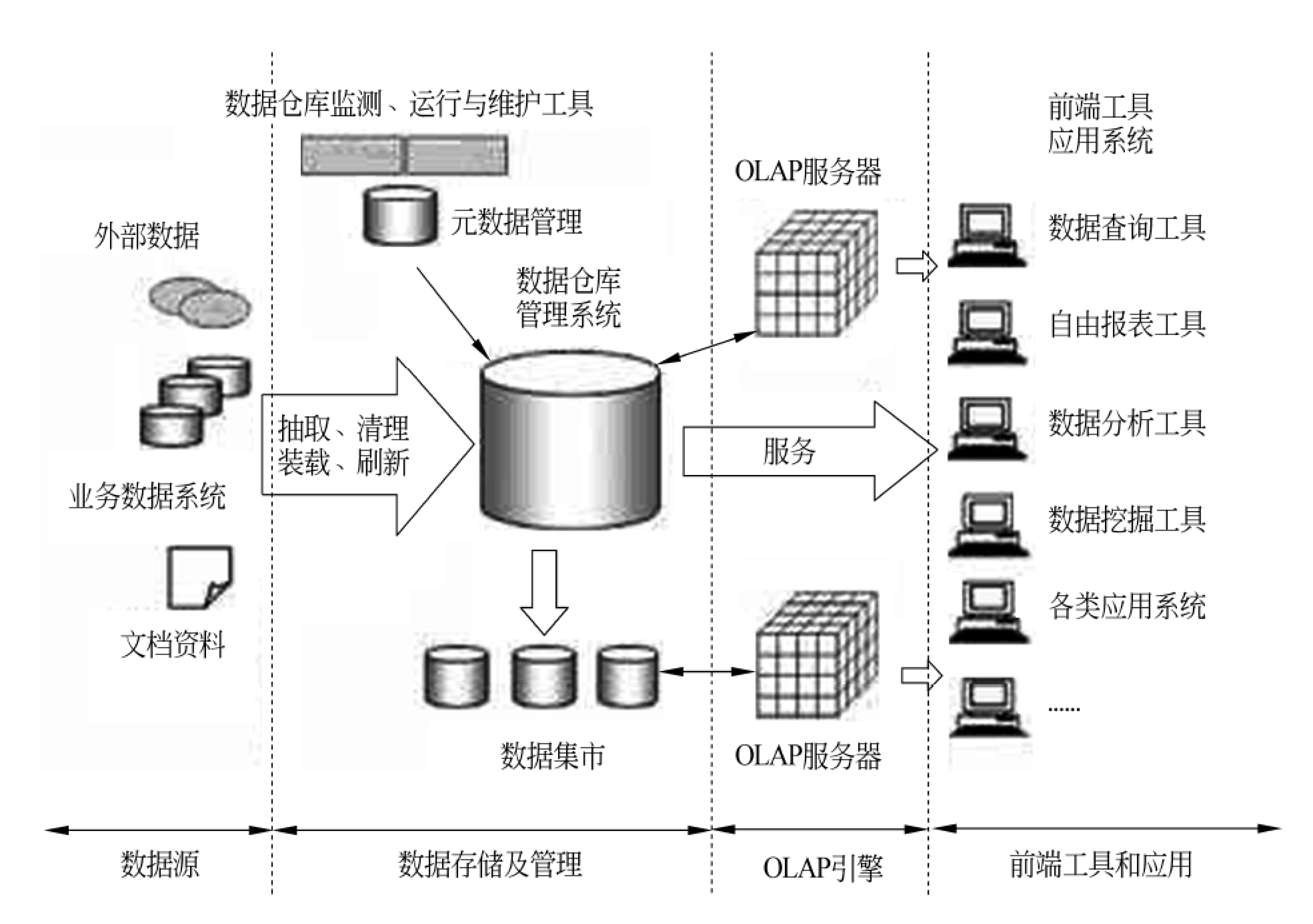


图 4.3 典型的数据仓库系统

2. 数据仓库系统的构成

一个典型的数据仓库系统主要有以下几部分构成：

1）数据仓库数据库

数据仓库数据库是整个数据仓库环境的核心，是数据存放的地方和提供对数据检索的支持。相对于操纵型数据库来说其突出的特点是对海量数据的支持和快速的检索技术。

2）数据抽取工具

数据抽取工具把数据从各种各样的存储方式中拿出来，进行必要的转化、整理，再存放到数据仓库内。对各种不同数据存储方式的访问能力是数据抽取工具的关键，应能生成 COBOL 程序、MVS 作业控制语言(JCL)、UNIX 脚本、和 SQL 语句等，以访问不同的数据。数据转换包括：删除对决策应用没有意义的数据段；转换到统一的数据名称和定义；计算统计和衍生数据；将默认值数据赋给默认值；把不同的数据定义方式统一。

3）元数据

元数据是描述数据仓库内数据的结构和建立方法的数据。可将其按用途的不同分为两类：技术元数据和商业元数据。

技术元数据是数据仓库的设计和管理人员用于开发和日常管理数据仓库使用的数据。包括：数据源信息；数据转换的描述；数据仓库内对象和数据结构的定义；数据清理和数据更新时用的规则；源数据到目的数据的映射；用户访问权限；数据备份历史记录、数据导入历史记录、信息发布历史记录等。

商业元数据从商业业务的角度描述了数据仓库中的数据。包括业务主题的描述以及所包含的数据、查询、报表。

元数据为访问数据仓库提供了一个信息目录(information directory),这个目录全面描述了数据仓库中都有什么数据、这些数据怎么得到的和怎么访问这些数据。它是数据仓库运行和维护的中心,数据仓库服务器利用它来存储和更新数据,用户通过它来了解和访问数据。

4) 访问工具

访问工具为用户访问数据仓库提供手段。有数据查询和报表工具、应用开发工具、管理信息系统工具、在线分析(OLAP)工具和数据挖掘(DM)工具等。

5) 数据集市(Data Marts)

数据集市是为了特定的应用目的或应用范围而从数据仓库中独立出来的一部分数据,也可称为部门数据或主题数据(subject area)。在数据仓库的实施过程中,往往可以从一个部门的数据集市着手,以后再用几个数据集市组成一个完整的数据仓库。需要注意的就是在实施不同的数据集市时,同一含义的字段定义一定要相容,这样才能保证以后实施数据仓库时不会造成大麻烦。

6) 数据仓库管理

数据仓库管理包括安全和特权管理、跟踪数据的更新、数据质量检查、管理和更新元数据、审计和报告数据仓库的使用和状态、删除数据、复制、分割和分发数据、备份和恢复以及存储管理。

7) 信息发布系统

信息发布系统的作用是把数据仓库中的数据或其他相关的数据发送给不同的地点或用户。基于 Web 的信息发布系统是对付多用户访问的最有效方法。

3. 数据仓库系统相关概念简介

1) 数据仓库数据库

以企业数据采集为目的,为了使得跨表或跨数据库(有时甚至是跨服务器)的汇总输出变得快速、高效率,而创建的一个可供数据分析查询用的信息中心储备库。这就是数据仓库数据库的含义。来自系统不同部分的信息被集成到数据仓库数据库中,以便于访问。

2) 联机事务处理(OLTP)

企业级关系数据库管理软件旨在集中存储由大公司或政府机构中的日常事务所产生的数据。由于这些系统基于计算机并记录企业的业务事务,因此被称为联机事务处理(OLTP) 系统。

3) 联机分析处理(OLAP)

联机分析处理是使分析人员、管理人员或执行人员能够从多种角度对从原始数据中转化出来的,能够真正为用户所理解的,并真实反映企业维特性的信息进行快速、一致、交互地存取,从而获得对数据的更深入了解的一类软件技术。

4) 数据挖掘(DM)

数据挖掘是指从大量原始数据中抽取模式的一个处理过程,抽取出来的模式就是所谓的知识,必须具备可信、新颖、有效和易于理解这四个特点。

4.6.4 数据仓库的基本架构

数据仓库的目的是构建面向分析的集成化数据环境，为企业提供决策支持(Decision Support)。其实数据仓库本身并不"生产"任何数据，同时自身也不需要"消费"任何的数据，数据来源于外部，并且开放给外部应用，这也是为什么叫"仓库"，而不叫"工厂"的原因。因此数据仓库的基本架构主要包含的是数据流入流出的过程，可以分为三层——源数据、数据仓库、数据应用，详见图 4.4。

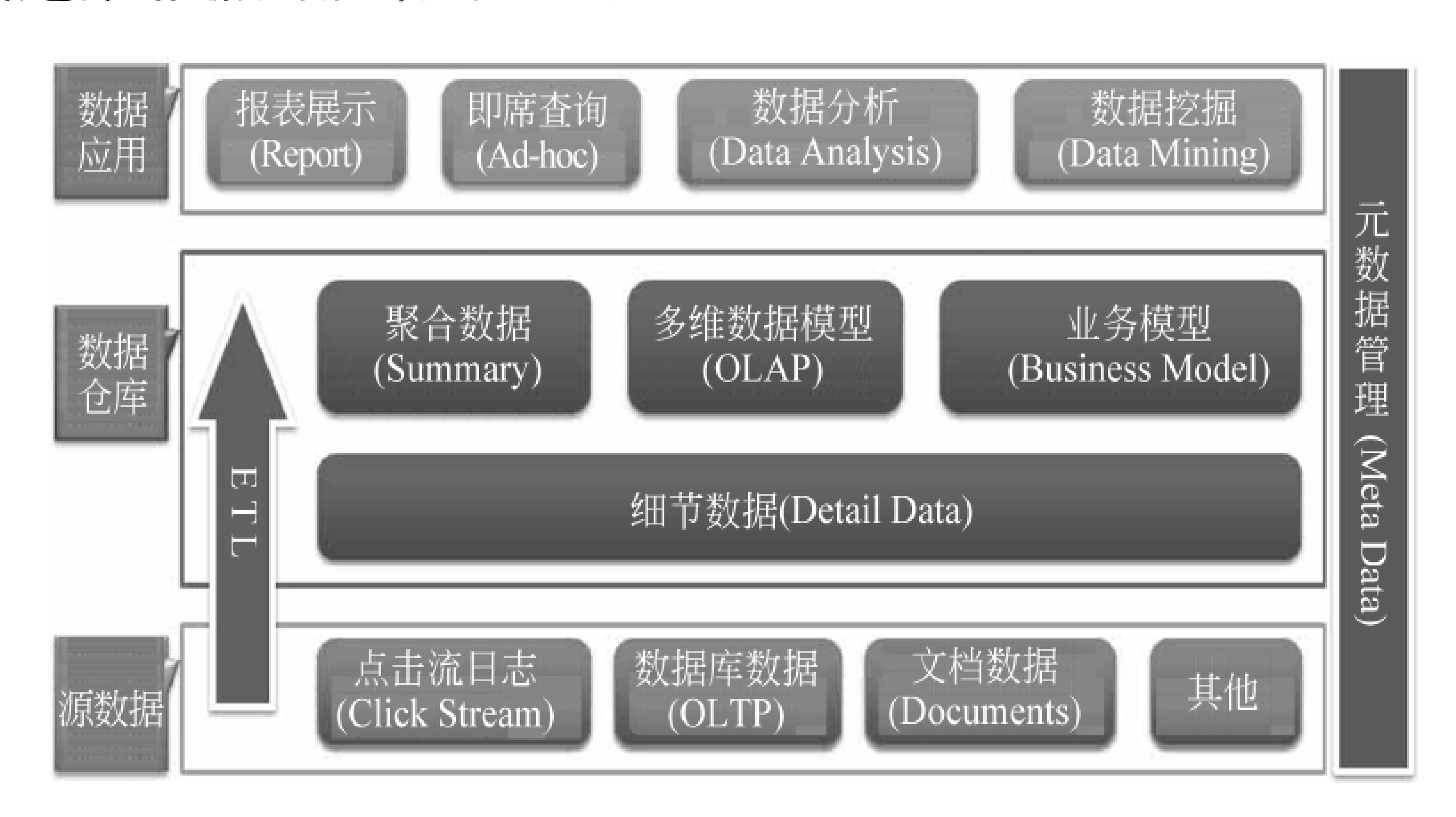


图 4.4　数据仓库的基本架构

从图 4.4 中可以看出数据仓库的数据来源于不同的源数据，并提供多样的数据应用，数据自上而下流入数据仓库后向上层开放应用，而数据仓库只是中间集成化数据管理的一个平台。

数据仓库从各数据源获取数据及在数据仓库内的数据转换和流动都可以认为是 ETL(抽取 Extra、转化 Transfer、装载 Load)的过程，ETL 是数据仓库的流水线，也可以认为是数据仓库的血液，它维系着数据仓库中数据的新陈代谢，而数据仓库日常的管理和维护工作的大部分精力就是保持 ETL 的正常和稳定。

下面主要简单介绍数据仓库架构中的各个模块，当然这里所介绍的数据仓库主要是指网站数据仓库。

4.6.5 数据仓库的数据存储

源数据通过 ETL 的日常任务调度导出，并经过转换后以特性的形式存入数据仓库。其实这个过程一直有很大的争议，就是到底数据仓库需不需要存储细节数据，一方的观点是数据仓库面向分析，所以只要存储特定需求的多维分析模型；另一方的观点是数据仓库先要建立和维护细节数据，再根据需求聚合和处理细节数据生成特定的分析模型。本书比较认同后面一个观点：数据仓库并不需要储存所有的原始数据，但数据仓库需要储存细节数据，并且导入的数据必须经过整理和转换使其面向主题。理由如下：

(1) 为什么不需要所有原始数据？

数据仓库面向分析处理，但是某些源数据对于分析而言没有价值或者其可能产生的

价值远低于存储这些数据所需要的数据仓库的实现和性能上的成本。比如我们知道用户的省份、城市足够，至于用户究竟住哪里可能只是物流商关心的事，或者用户在博客的评论内容可能只是文本挖掘会有需要，但将这些冗长的评论文本存在数据仓库就得不偿失。

(2) 为什么要保存细节数据？

细节数据是必需的，数据仓库的分析需求会时刻变化，而有了细节数据就可以做到以不变应万变，但如果我们只存储根据某些需求搭建起来的数据模型，那么显然对于频繁变动的需求会手足无措。

(3) 为什么要面向主题？

面向主题是数据仓库的第一特性，主要是指合理地组织数据以方面实现分析。对于源数据而言，其数据组织形式是多样的，像点击流的数据格式是未经优化的，前台数据库的数据是基于 OLTP 操作组织优化的，这些可能都不适合分析，而整理成面向主题的组织形式才是真正有利于分析的，比如将点击流日志整理成页面（Page）、访问（Visit 或 Session）、用户（Visitor）三个主题，这样可以明显提升分析的效率。

数据仓库基于维护细节数据的基础上在对数据进行处理，使其能够真正地应用于分析。主要包括三个方面：

(1) 数据的聚合。

这里的聚合数据指的是基于特定需求的简单聚合（基于多维数据的聚合体现在多维数据模型中），简单聚合可以是网站的总 Pageviews、Visits、Unique Visitors 等汇总数据，也可以是 Avg. time on page、Avg. time on site 等平均数据，这些数据可以直接地展示于报表上。

(2) 多维数据模型。

多维数据模型提供了多角度多层次的分析应用，比如基于时间维、地域维等构建的销售星形模型、雪花模型，可以实现在各时间维度和地域维度的交叉查询，以及基于时间维和地域维的细分。所以多维数据模型的应用一般都是基于联机分析处理（Online Analytical Process，OLAP）的，而面向特定需求群体的数据集市也会基于多维数据模型进行构建。

(3) 业务模型。

这里的业务模型指的是基于某些数据分析和决策支持而建立起来的数据模型，比如用户评价模型、关联推荐模型、RFM 分析模型等，或者是决策支持的线性规划模型、库存模型等；同时，数据挖掘中前期数据的处理也可以在这里完成。

4.6.6　数据仓库的数据应用

以上介绍了数据仓库的四大特性上的价值体现，但数据仓库的价值远不止这些，而且其价值真正的体现是在数据仓库的数据应用上，一切数据相关的扩展性应用都可以基于数据仓库来实现。

1. 报表展示

报表几乎是每个数据仓库的必不可少的一类数据应用，将聚合数据和多维分析数据

展示到报表，提供了最为简单和直观的数据。

2. 即席查询

理论上数据仓库的所有数据（包括细节数据、聚合数据、多维数据和分析数据）都应该开放即席查询，即席查询提供了足够灵活的数据获取方式，用户可以根据自己的需要查询获取数据，并提供导出到 Excel 等外部文件的功能。

3. 数据分析

数据分析大部分可以基于构建的业务模型展开，当然也可以使用聚合的数据进行趋势分析、比较分析、相关分析等，而多维数据模型提供了多维分析的数据基础；同时从细节数据中获取一些样本数据进行特定的分析也是较为常见的一种途径。

4. 数据挖掘

数据挖掘用一些高级的算法可以让数据展现出各种令人惊讶的结果。数据挖掘可以基于数据仓库中已经构建起来的业务模型展开，但大多数时候数据挖掘会直接从细节数据上入手，而数据仓库为挖掘工具诸如 SAS、SPSS 等提供数据接口。

4.6.7 元数据管理

元数据（Meta Date）其实应该叫做解释性数据，即数据的数据。主要记录数据仓库中模型的定义、各层级间的映射关系、监控数据仓库的数据状态及 ETL 的任务运行状态。一般会通过元数据资料库（Metadata Repository）来统一地存储和管理元数据，其主要目的是使数据仓库的设计、部署、操作和管理能达成协同和一致。

最后做个结论，数据仓库本身既不生产数据也不消费数据，只是作为一个中间平台集成化地存储数据；数据仓库实现的难度在于整体架构的构建及 ETL 的设计，这也是日常管理维护中的重头；而数据仓库的真正价值体现在基于它的数据应用上，如果没有有效的数据应用，也就失去了构建数据仓库的意义。

4.7 大数据应用案例之：一场雾霾将损失多少 GDP

2005—2010 年，全球因空气污染的死亡率上升了 4%，其中，中国上升了 5%。2010 年，北京、上海、广州、西安四城市因 PM2.5 污染造成 7770 人早死。

正当全球领导人汇聚巴黎讨论气候问题之时，中国大面积遭遇严重雾霾污染（见图 4.5）。一场切肤之痛，再次引发人们对环境问题的关注，也生发出更多质问和治理思考。

在无法逃避的情况下，人们只能自嘲或他嘲，齐齐做出“等风来”的祈祷状。但必须指出的是，空气污染问题比你想象的更为严重。联合国环境规划署早已将空气污染列为“全球最严重的环境健康风险”。

其中，直径在 2.5μm 及以内的细微颗粒物（PM2.5）产生于化石燃料和生物质的不完全燃烧，是人们最担心的空气污染问题之一。PM2.5 的直径是人的头发丝厚度的百分之一，它可以深入渗透到肺部和血液中，并且不论在何种浓度都是危险的。国际癌症研究机



图 4.5　中国大面积遭遇严重雾霾污染

构(IARC)在 2013 年确认颗粒物是人类致癌物。

2001—2006 年全球 PM2.5 浓度的统计分析图如图 4.6 所示。

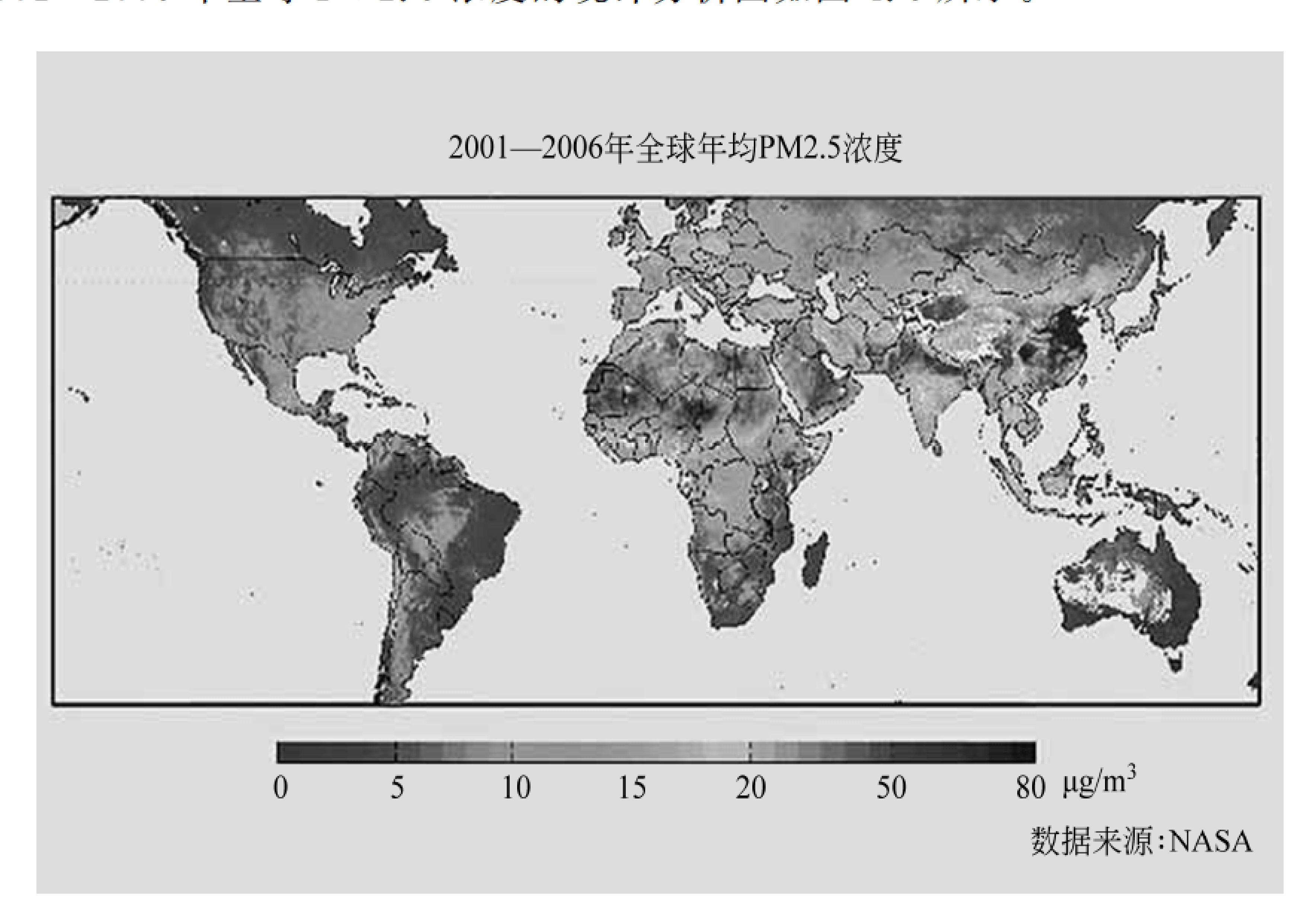


图 4.6　全球 PM2.5 浓度

1. 问题的严重性

那么,这位"杀手"到底有多厉害呢?

世界卫生组织在给当地机关的留言中披露,全世界只有 12%的城市达到了世界卫生组织空气污染指导标准,许多城市市中心的污染达到了建议水平的 10 倍以上。

而据世界卫生组织给各国财政部长的留言信,从全球来看,能源消费最大的负外部性就是空气污染,每年造成超过 700 万人死亡。这是之前预估的两倍还多。其中,超过 400 万人死于室内空气污染,超过 350 万人死于室外空气污染。室外空气污染致死人数统计分析见图 4.7 所示。

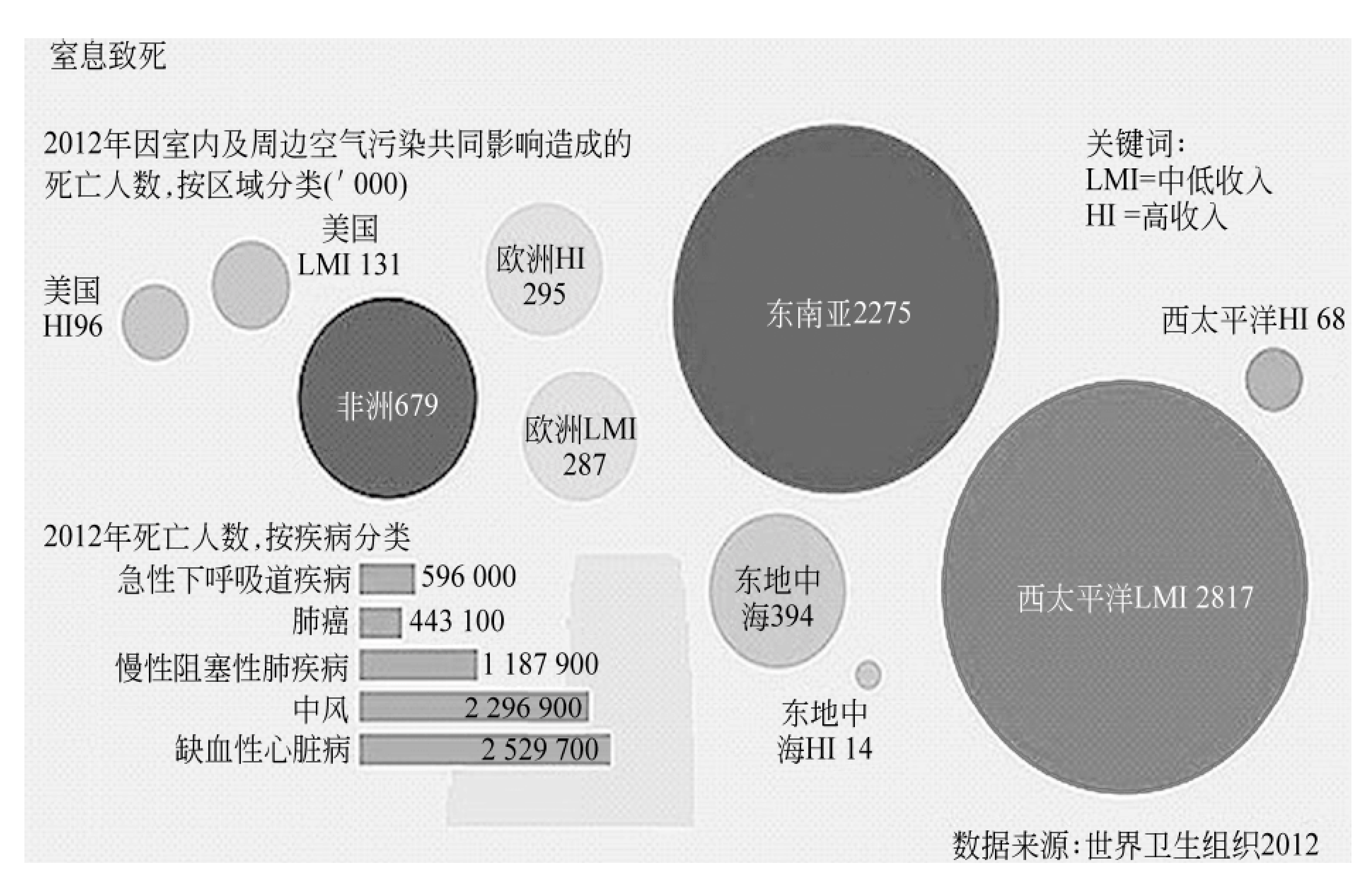


图 4.7 室外空气污染致死人数

而不幸的是,这种死亡威胁仍在加强。联合国数据称,从 2005—2010 年,全球死亡率上升了 4%,其中,中国上升了 5%,印度上升了 12%。

绿色和平组织和北京大学公共卫生学院于 2012 年底共同发布的《危险的呼吸——PM2.5 的健康危害和经济损失评估研究》指出,空气污染致死已被研究证实,2010 年,北京、上海、广州、西安四城市因 PM2.5 污染造成 7770 人早死。

不仅如此,《联合国环境规划署年鉴 2014》显示,在大多数室外空气污染受监测的城市,其空气质量都达不到世界卫生组织指南中关于可接受污染水平的标准。在这些城市生活的居民拥有更高的患上中风、心脏病、肺癌、慢性和急性呼吸道疾病(包括哮喘)及其他健康疾病的风险。

2. 经济损失

空气污染不仅侵蚀着人们的生命安全,也消耗着经济增长前景。

据经合组织(OECD)研究,仅仅 2010 年,空气污染给中国和印度造成的经济损失,就分别高达 1.4 万亿美元和 0.5 万亿美元。在欧洲,由于暴露于道路交通造成的空气污染中而导致的损失为每年 1370 亿美元,而 2009 年,由于 10 000 个污染设备产生的空气污染所造成的损失——人口死亡、疾病和作物损毁——约为 1400～2300 亿美元。

世界卫生组织在给各国财政部长的留言信中则表示,2015 年仅能源消费引起的室外空气污染一项,其造成的非补贴健康影响价值就达到了约 27 000 亿美元,超过了给能源部门支持总额的一半。

联合国环境规划署也预估,到 2030 年,全球由于地面臭氧污染造成的大豆、玉米、小麦等作物的损失可达每年 170～350 亿美元。

据联合国环境规划署数据,空气污染对世界最先进经济体,以及印度和中国造成的损失的估值已达到每年 3.5 万亿美元。这些损失主要是人口死亡和疾病问题。据估计,

2010 年，室外空气污染在经合组织国家(OECD)造成的人口死亡及疾病问题的经济影响为 1.7 万亿美元。

3. 收益

正因为空气污染问题很严重，更激发了全球为之寻求解决方案的努力，而且，改善空气质量带来的巨大经济效益潜力也显而易见。

在美国，由于 1990 年《清洁空气法案修正案》的实施而减少的 PM2.5 及地面臭氧，它们所带来的直接经济效益据估计是推行这项政策所使用经费的 90 倍。大约 85%的经济效益归因于户外环境中的 PM2.5 含量降低，从而使得过早死亡数量减少，仅在 2020 年一年就可以避免 23 万个过早死亡病例。

据世界卫生组织估计，一旦实施了国家性能源适宜定价，室外空气污染造成的死亡人数将减少三分之一，并能降低超过 20%的温室气体排放。

而且，如果取消能源补贴，转而改为设置与国家利益相一致的税收项目，那么，就能提升大约 3%的国内生产总值，相当于每年新增 3 万亿美元。世界卫生组织在给各国财政部长的留言中披露这一数据。

实际上，近年来空气污染稍有减缓，这部分归因于更严格的排放控制。世界银行的研究即发现，在撒哈拉沙漠以南的非洲，低硫燃料(50ppm)及清洁型交通工具(包括摩托车)的应用，预计将在十年的时间内产生 430 亿美元的健康收益。

(注：上述数据均据《联合国环境规划署年鉴 2014》、世界卫生组织网站)

习题与思考题

一、选择题

1. 大数据应用需依托的新技术有(　　)。

A. 大规模存储与计算　　B. 数据分析处理

C. 智能化　　D. 三个选项都是

2. 在数据生命周期管理实践中，(　　)是执行方法。

A. 数据存储和备份规范　　B. 数据管理和维护

C. 数据价值发觉和利用　　D. 数据应用开发和管理

3. 下列关于计算机存储容量单位的说法中，错误的是(　　)。

A. 1KB<1MB<1GB　　B. 基本单位是字节(Byte)

C. 一个汉字需要一个字节的存储空间　　D. 一个字节能够容纳一个英文字符

4. 数据仓库的最终目的是(　　)。

A. 收集业务需求　　B. 建立数据仓库逻辑模型

C. 开发数据仓库的应用分析　　D. 为用户和业务部门提供决策支持

5. 下列说法正确的是(　　)。

A. 有价值的数据是附属于企业经营核心业务的一部分数据

B. 数据挖掘它的主要价值后就没有必要再进行分析了

C. 所有数据都是有价值的

D. 在大数据时代,收集、存储和分析数据非常简单

6. 关于数据创新包含(　　)。(多选题)

A. 数据的再利用　　B. 重组数据

C. 可扩展数据　　D. 数据的折旧值

E. 数据废气　　F. 开放数据

7. 相比依赖于小数据和精确性的时代,大数据因为更强调数据的(　　),帮助我们进一步接近事实的真相。

A. 安全性　　B. 完整性　　C. 混杂性　　D. 完整性和混杂性

二、问答题

1. 传统数据存储有哪几种存储的模式?请简要说明。
2. 什么是分布式存储系统?什么是云存储?
3. 什么是 NoSQL 非结构化数据库?什么是大规模并行分析数据库?
4. 简述数据仓库原理及构成。
5. 简要说明数据仓库的基本架构。

第 5 章　大数据计算模式与处理系统

计算模式的出现有力推动了大数据技术和应用的发展，使其成为目前大数据处理最为成功、最广为接受使用的主流大数据计算模式。然而，现实世界中的大数据处理问题复杂多样，难以有一种单一的计算模式涵盖所有不同的大数据计算需求。

研究和实际应用中发现，由于 MapReduce 主要适合于进行大数据线下批处理，在面向低延迟和具有复杂数据关系和复杂计算的大数据问题时有很大的不适应性。因此，近几年来学术界和业界在不断研究并推出多种不同的大数据计算模式。

所谓大数据计算模式，即根据大数据的不同数据特征和计算特征，从多样性的大数据计算问题和需求中提炼并建立的各种高层抽象（Abstraction）或模型（Model）。

传统的并行计算方法主要从体系结构和编程语言的层面定义了一些较为底层的并行计算抽象和模型，但由于大数据处理问题具有很多高层的数据特征和计算特征，因此大数据处理需要更多地结合这些高层特征考虑更为高层的计算模式。

5.1　数据计算

面向大数据处理的数据查询、统计、分析、挖掘等需求，促生了大数据计算的不同计算模式，整体上我们把大数据计算分为离线批处理计算、实时交互计算和流计算三种。

5.1.1　离线批处理

随着云计算技术到广泛的应用的发展，基于开源的 Hadoop 分布式存储系统和 MapReduce 数据处理模式的分析系统也得到了广泛的应用。

Hadoop 通过数据分块及自恢复机制，能支持 PB 级的分布式的数据存储，以及基于 MapReduce 分布式处理模式对这些数据进行分析和处理。MapReduce 编程模型可以很容易地将多个通用批数据处理任务和操作在大规模集群上并行化，而且有自动化的故障转移功能。MapReduce 编程模型在 Hadoop 这样的开源软件带动下被广泛采用，应用到 Web 搜索、欺诈检测等各种各样的实际应用中。

Hadoop 是一个能够对大量数据进行分布式处理的软件框架，而且是以一种可靠、高效、可伸缩的方式进行处理，依靠横向扩展，通过不断增加廉价的商用服务器来提高计算和存储能力。用户可以轻松地在上面开发和运行处理海量数据的应用程序。以 Hadoop 平台为代表的大数据处理平台技术包括 MapReduce、HDFS、HBase、Hive、Zookeeper、Avro 和 Pig 等，已经形成了一个 Hadoop 生态圈，如图 5.1 所示。

MapReduce 编程模型是 Hadoop 的心脏，用于大规模数据集的并行运算。正是这种

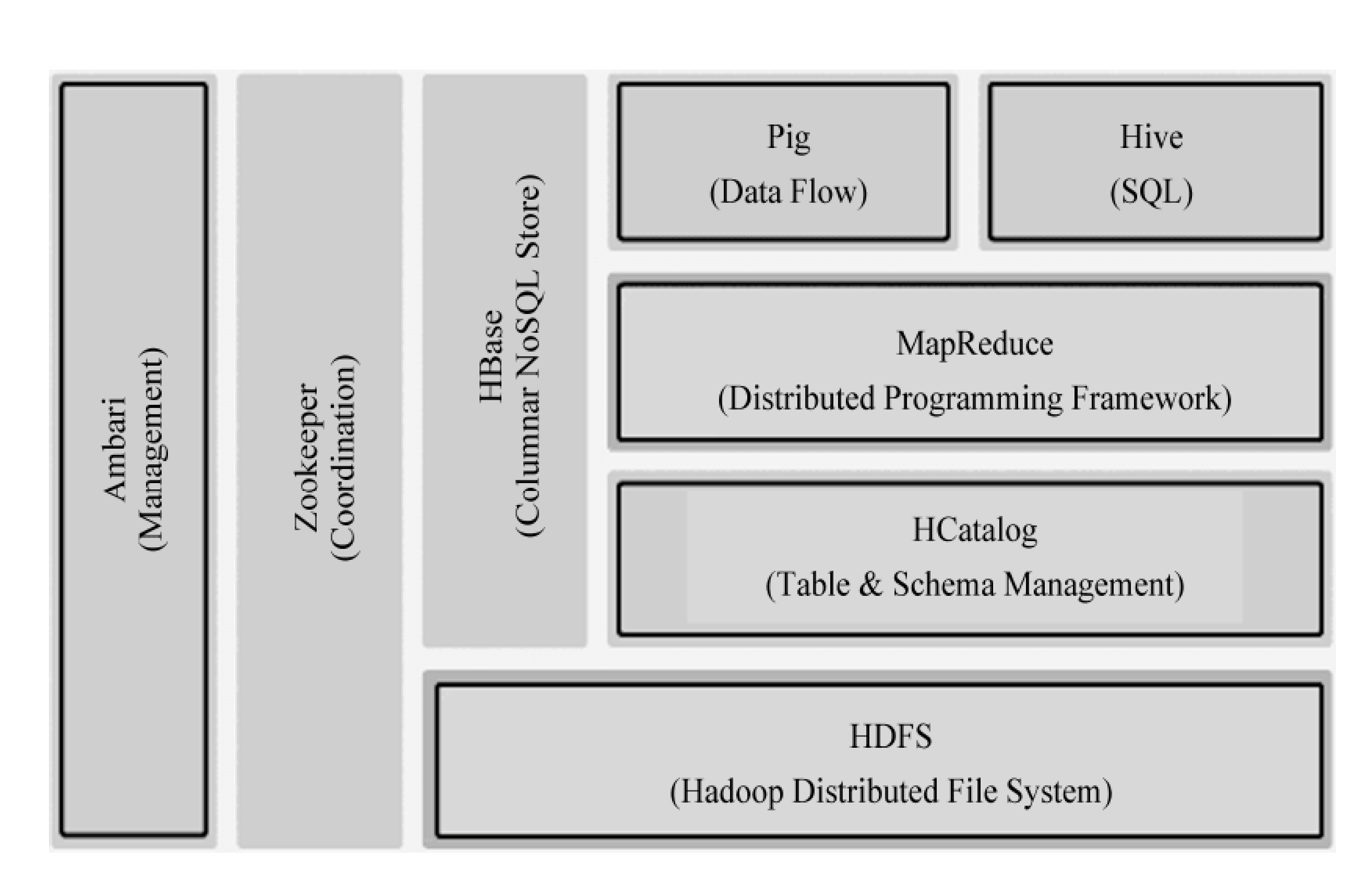


图 5.1 Hadoop 生态圈

编程模式，实现了跨越一个 Hadoop 集群中数百或数千台服务器的大规模扩展性。

分布式文件系统 HDFS 提供基于 Hadoop 处理平台的海量数据存储，其中的 NameNode 提供元数据服务，DataNode 用于存储文件系统的文件块。

HBase 是建立在 HDFS 之上，用于提供高可靠性、高性能、列存储、可伸缩、实时读写的数据库系统，可以存储非结构化和半结构化的松散数据。

Hive 是基于 Hadoop 的大型数据仓库，可以用来进行数据的提取、转化和加载(ETL)，存储、查询和分析存储在 Hadoop 中的大规模数据。

Pig 是基于 Hadoop 的大规模数据分析平台，可以把类 SQL 的数据分析请求转换为一系列经过优化处理的 MapReduce 运算，为复杂的海量数据并行计算提供了一个简单的操作和编程接口。

Zookeeper 是高效、可靠的协同工作系统，用于协调分布式应用上的各种服务，利用 Zookeeper 可以构建一个有效防止单点失效及处理负载均衡的协调服务。

Avro 作为二进制的高性能的通信中间件，提供了 Hadoop 平台间的数据序列化功能和 RPC 服务。

但 Hadoop 平台主要是面向离线批处理应用的，典型的是通过调度批量任务操作静态数据，计算过程相对缓慢，有的查询可能会花几小时甚至更长时间才能产生结果，对于实时性要求更高的应用和服务则显得力不从心。

MapReduce 是一种很好的集群并行编程模型，能够满足大部分应用的需求。虽然 MapReduce 是分布式/并行计算方面一个很好的抽象，但它并不一定适合解决计算领域的任何问题。例如，对于那些需要实时获取计算结果的应用，像基于流量的点击付费模式的广告投放、基于实时用户行为数据分析的社交推荐、基于网页检索和点击流量的反作弊统计等等。对于这些实时应用，MapReduce 并不能提供高效处理，因为处理这些应用逻辑需要执行多轮作业，或者需要将输入数据的粒度切分到很小。

5.1.2　实时交互计算

当今的实时计算一般都需要针对海量数据进行，除了要满足非实时计算的一些需求（如计算结果准确）以外，实时计算最重要的一个需求是能够实时响应计算结果，一般要求为秒级。实时计算一般可以分为以下两种应用场景：

（1）数据量巨大且不能提前计算出结果的，但要求对用户的响应时间是实时的。

主要用于特定场合下的数据分析处理.当数据量庞大，同时发现无法穷举所有可能条件的查询组合，或者大量穷举出来的条件组合无用的时候，实时计算就可以发挥作用，将计算过程推迟到查询阶段进行，但需要为用户提供实时响应。这种情形下，也可以将一部分数据提前进行处理，再结合实时计算结果，以提高处理效率。

（2）数据源是实时的和不间断的，要求对用户的响应时间也是实时的。

数据源实时不间断的也称为流式数据。所谓流式数据，是指将数据看作是数据流的形式来处理。数据流是在时间分布和数量上无限的一系列数据记录的集合体；数据记录是数据流的最小组成单元。例如，在物联网领域传感器产生的数据可能是源源不断的，实时的数据计算和分析可以动态实时地对数据进行分析统计，对于系统的状态监控、调度管理具有重要的实际意义。

5.1.3　海量数据实时计算

海量数据的实时计算过程可以被划分为以下三个阶段：数据的产生与收集阶段、传输与分析处理阶段、存储和对外提供服务阶段，如图5.2所示。

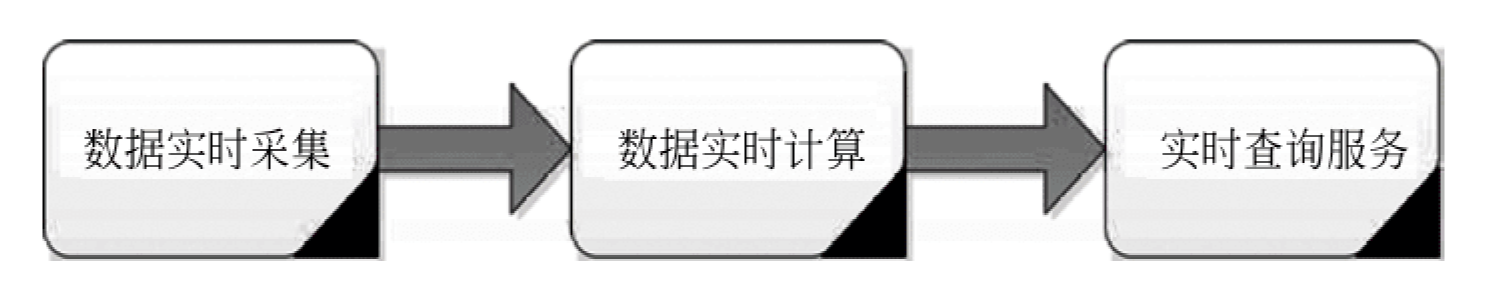


图5.2　实时计算过程

1. 数据实时采集

数据实时采集在功能上需要保证可以完整地收集到所有数据，为实时应用提供实时数据；响应时间上要保证实时性、低延迟；配置简单，部署容易；系统稳定可靠等。目前，互联网企业的海量数据采集工具包括Facebook开源的Scribe、LinkedIn开源的Kafka、Cloudera开源的Flume、淘宝开源的TimeTunnel、Hadoop的Chukwa等，均可以满足每秒数百MB的日志数据采集和传输需求。

2. 数据实时计算

传统的数据操作，首先将数据采集并存储在数据库管理系统（DBMS）中，然后通过query和DBMS进行交互，得到用户想要的答案。整个过程中，用户是主动的，而DBMS系统是被动的。但是，对于现在大量存在的实时数据，这类数据实时性强、数据量大、数据格式多种多样，传统的关系型数据库架构并不合适。新型的实时计算架构一般都是采用海量并行处理MPP的分布式架构，数据的存储及处理会分配到大规模的结点上进行，以

满足实时性要求，在数据的存储上，则采用大规模分布式文件系统，比如，Hadoop 的 HDFS 文件系统，或是新型的 NoSQL 分布式数据库。

3. 实时查询服务

实时查询服务的实现可以分为三种方式。

(1) 全内存：直接提供数据读取服务，定期 dump 到磁盘或数据库进行持久化。

(2) 半内存：使用 Redis、Memcache、MongoDB、BerkeleyDB 等数据库提供数据实时查询服务，由这些系统进行持久化操作。

(3) 全磁盘：使用 HBase 等以分布式文件系统(HDFS)为基础的 NoSQL 数据库，对于 Key-Value 引擎，关键是设计好 Key 的分布。

实时和交互式计算技术中，Google 的 Dremel 系统表现最为突出。Dremel 是 Google 的"交互式"数据分析系统，可以组建成规模上千的集群，处理 PB 级别的数据。作为 MapReduce 的发起人，Google 开发了 Dremel 系统将处理时间缩短到秒级，作为 MapReduce 的有力补充。

Dremel 作为 Google BigQuery 的 report 引擎，获得了很大的成功。与 MapReduce 一样，Dremel 也需要和数据运行在一起，将计算移动到数据上面。它需要 GFS 这样的文件系统作为存储层。Dremel 支持一个嵌套(nested)的数据模型，类似于 JSON。而传统的关系模型由于不可避免地有大量的 Join 操作，在处理如此大规模的数据的时候，往往是有心无力。Dremel 同时还使用列式存储，分析的时候，可以只扫描需要的那部分数据，以减少 CPU 和磁盘的访问量。同时列式存储是压缩友好的，使用压缩，可以减少存储量，发挥最大的效能。

5.1.4 流计算

在很多实时应用场景中，比如实时交易系统、实时诈骗分析、实时广告推送、实时监控、社交网络实时分析等，数据量大，实时性要求高，而且数据源是实时不间断的。新到的数据必须马上处理完，不然后续的数据就会堆积起来，永远也处理不完。反应时间经常要求在秒级以下，甚至是毫秒级，这就需要一个高度可扩展的流式计算解决方案。

流计算就是针对实时连续的数据类型而准备的。在流数据不断变化的运动过程中实时地进行分析，捕捉到可能对用户有用的信息，并把结果发送出去。在整个过程中，数据分析处理系统是主动的，用户处于被动接收的状态，如图 5.3 所示。

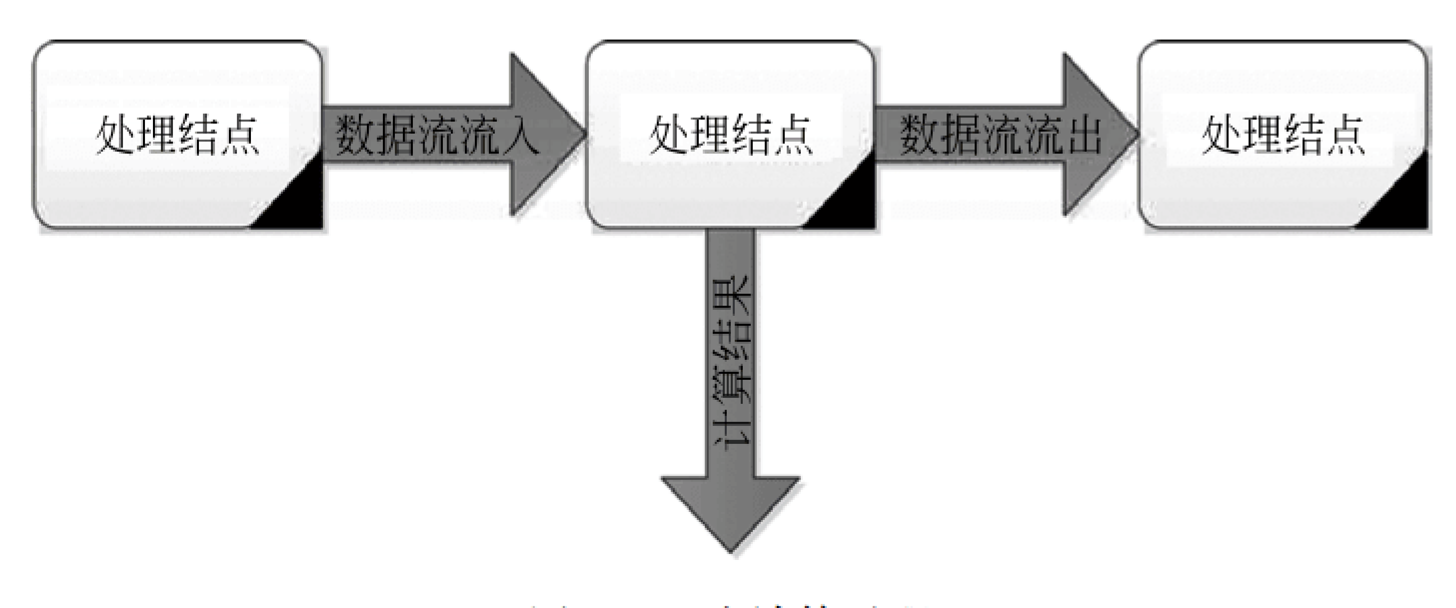


图 5.3 流计算过程

传统的流式计算系统，一般是基于事件机制，所处理的数据量也不大。新型的流处理技术，如 Yahoo 的 S4 主要解决的是高数据率和大数据量的流式处理。

S4 是一个通用的、分布式的、可扩展的、部分容错的、可插拔的平台。开发者可以很容易地在其上开发面向无界不间断流数据处理的应用。

5.2　聚类算法

聚类分析是一种重要的人类行为，早在孩提时代，一个人就通过不断改进下意识中的聚类模式来学会如何区分猫狗、动物植物。目前在许多领域都得到了广泛的研究和成功的应用，如用于模式识别、数据分析、图像处理、市场研究、客户分割、Web 文档分类等。

聚类就是按照某个特定标准(如距离准则)把一个数据集分割成不同的类或簇，使得同一个簇内的数据对象的相似性尽可能大，同时不在同一个簇中的数据对象的差异性也尽可能地大。即聚类后同一类的数据尽可能聚集到一起，不同数据尽量分离。

聚类技术正在蓬勃发展，对此有贡献的研究领域包括数据挖掘、统计学、机器学习、空间数据库技术、生物学以及市场营销等。各种聚类方法也被不断提出和改进，而不同的方法适合于不同类型的数据，因此对各种聚类方法、聚类效果的比较成为值得研究的课题。

5.2.1　聚类算法的分类

目前，有大量的聚类算法。而对于具体应用，聚类算法的选择取决于数据的类型、聚类的目的。如果聚类分析被用作描述或探查的工具，可以对同样的数据尝试多种算法，以发现数据可能揭示的结果。

主要的聚类算法可以划分为如下几类：划分方法、层次方法、基于密度的方法、基于网格的方法以及基于模型的方法。

每一类中都存在着得到广泛应用的算法，例如，划分方法中的 k-mean 聚类算法、层次方法中的凝聚型层次聚类算法、基于模型方法中的神经网络聚类算法等。

目前，聚类问题的研究不仅仅局限于上述的硬聚类，即每一个数据只能被归为一类，模糊聚类也是聚类分析中研究较为广泛的一个分支。模糊聚类通过隶属函数来确定每个数据隶属于各个簇的程度，而不是将一个数据对象硬性归类到某一簇中。目前已有很多关于模糊聚类的算法被提出，如著名的 FCM 算法等。

5.2.2　数据分类与聚类

聚类的算法有很多，现在已知的算法主要有四种类型：划分聚类、层次聚类、基于密度的聚类、基于表格的聚类。

1. 划分聚类

对于给定的数据集，划分聚类需要知道要划分簇的数目 k($k \leqslant n$，n 是数据集中项的数目)。划分聚类将数据分为 k 组，每组至少有一项。大多数划分聚类都是基于距离的。一般情况下给出了聚类数目 k，首先会产生一个初始的划分，然后用迭代的方法通过更改数

据项所属的簇来提高划分的质量。一个好的划分的标准是同一个簇内的数据项彼此相似，相反地，不同簇的项有较大的区别。

实现全局最优划分往往很难在复杂度忍受的范围内做到。然而，大多数应用都选取了一些启发式方法。比如像选取贪心策略的 *k*-means 和 *k*-medoids 算法，都极大地提高了划分质量，并达到了一个局部最优解。这些启发式聚类算法在中小型数据集中挖掘类似球形簇表现非常好。

2. 层次聚类

层次聚类就是通过对数据集按照某种方法进行层次分解，直到满足某种条件为止。层次聚类根据划分的方法分为凝聚和分割两种。凝聚的方法也叫做自底向上方法。它每次迭代将最相近两个项（或者组）合并形成一个新的组，直至最终形成一个组或者达到其他停止的条件。

分割的方法也叫自顶向下，与凝聚的方法相反。开始的时候讲所有数据看成一个组，每一次迭代一个簇就被划分成两个小一点儿的簇。直到最终每个项都是一个簇或者达到了某个停止条件。层次聚类可以是基于距离、基于密度、基于连接的。层次聚类有一个缺点：一旦一个凝聚或分割形成了，这个操作就永远不能再更改了。这样的好处就是计算复杂度相对较低。

3. 基于密度的聚类

很多聚类算法都是根据距离计算的。这样很容易发现球形的簇，很难发现其他形状的簇。基于密度的算法认为，在整个样本空间点中，各目标类簇是由一群的稠密样本点组成的，而这些稠密样本点被低密度区域（噪声）分割，而算法的目的就是要过滤低密度区域，发现稠密样本点。这类算法往往重视数据项的密集程度，因此这些算法都是基于连接的。虽然是基于连接的，但是也强调了连接过程中数据项周围的密度。这样就能发现各种任意形状的聚类簇。

4. 基于网格的聚类

这类算法将数据项的空间划分成有限数目的网格。所有的聚类操作都是在网格上进行的。这样最大的好处是计算速度相当快。因为计算过程跟数据项的数目没有关系，只与每一维网格的数目和维数有关系。对于大数据的数据挖掘问题，网格的方法效率往往会很不错。然而网格只是一种思想，这种思想往往要和其他的算法相结合才能解决好实际问题，比如聚类。

5.3 数据集成

近几十年来，科学技术的迅猛发展和信息化的推进，使得人类社会所积累的数据量已经超过了过去 5000 年的总和，数据的采集、存储、处理和传播的数量也与日俱增。企业实现数据共享，可以使更多的人更充分地使用已有的数据资源，减少资料收集、数据采集等重复劳动和相应费用。

但是，在实施数据共享的过程当中，由于不同用户提供的数据可能来自不同的途径，

其数据内容、数据格式和数据质量千差万别，有时甚至会遇到数据格式不能转换或数据转换格式后丢失信息等棘手问题，严重阻碍了数据在各部门和各软件系统中的流动与共享。因此，如何对数据进行有效的集成管理已成为增强企业商业竞争力的必然选择。

由于现代企业的飞速发展和企业逐渐从一个孤立结点发展成为不断与网络交换信息和进行商务事务的实体，企业数据交换也从企业内部走向了企业之间；同时，数据的不确定性和频繁变动，以及这些集成系统在实现技术和物理数据上的紧耦合关系，导致一旦应用发生变化或物理数据变动，整个体系将不得不随之修改。因此，我们进行数据集成将面临如何适应现代社会发展的复杂需求、有效扩展应用领域、分离实现技术和应用需求、充分描述各种数据源格式以及发布和进行数据交换等问题。

5.3.1　数据集成概述

1. 数据集成模型分类

数据集成是把不同来源、格式、特点、性质的数据在逻辑上或物理上有机地集中，从而为企业提供全面的数据共享。在企业数据集成领域，已经有了很多成熟的框架可以利用。目前通常采用联邦式、基于中间件模型和数据仓库等方法来构造集成的系统，这些技术在不同的着重点和应用上解决数据共享和为企业提供决策支持。在这里将对这几种数据集成模型做一个基本的分析。

1）联邦数据库系统（FDBS）

由半自治数据库系统构成，相互之间分享数据，联盟各数据源之间相互提供访问接口，同时联盟数据库系统可以是集中数据库系统或分布式数据库系统及其他类型数据库，松耦合而不提供统一的接口，但可以通过统一的语言访问数据源，其中的核心是必须解决所有数据源语义上的问题。

2）中间件模式

是目前比较流行的数据集成方法，它通过在中间层提供一个统一的数据逻辑视图来隐藏底层的数据细节，使得用户可以把集成数据源看为一个统一的整体。这种模型下的关键问题是如何构造这个逻辑视图并使得不同数据源之间能映射到这个中间层。

通过统一的全局数据模型来访问异构的数据库、遗留系统、Web 资源等。中间件位于异构数据源系统（数据层）和应用程序（应用层）之间，向下协调各数据源系统，向上为访问集成数据的应用提供统一数据模式和数据访问的通用接口。各数据源的应用仍然完成它们的任务，中间件系统则主要集中为异构数据源提供一个高层次检索服务。

3）数据仓库

数据仓库是在企业管理和决策中面向主题的、集成的、与时间相关的和不可修改的数据集合。其中，数据被归类为广义的、功能上独立的、没有重叠的主题。这几种方法在一定程度上解决了应用之间的数据共享和互通的问题，但也存在以下的异同：联邦数据库系统主要面向多个数据库系统的集成，其中数据源有可能要映射到每一个数据模式，当集成的系统很大时，对实际开发将带来巨大的困难。

数据仓库技术在另外一个层面上表达数据之间的共享，它主要是为了针对企业某个应用领域提出的一种数据集成方法，也就是我们在上面所提到的面向主题并为企业提供

数据挖掘和决策支持的系统。

2. 数据高速缓存器是关键

对数据集成体系结构来说，关键是拥有一个包含有目标计划、源-目标映射、数据获得、分级抽取、错误恢复和安全性转换的数据高速缓存器。此外，数据高速缓存器包含有预先定制的数据抽取工作，这些工作自动位于一个企业的后端及数据仓库之中。

一个高速缓存器作为企业和电子商务数据的一个单一集成点，最大限度地减少了对直接访问后端系统和进行复杂实时集成的需求。这个高速缓存器从后端系统中卸载众多不必要的数据请求，因此使电子商务公司可以增加更多的用户，同时让后端系统从事其指定的工作。

数据集成软件与企业应用集成厂商和程序集成商进行联合，而不是取代它们。的确，由于数据集成软件越来越普遍地被用来作为 B2B 集成的一个工具，它会引人注目地改造 B2B 集成商一起合作的方式以及企业向 Internet 迁移的方式。

3. 数据集成对于企业信息系统的作用

数据集成的出现使企业能够将后端的 ERP 信息迁移到 Internet 上。数据集成产品在一个公司的 Internet 计算机与 SAP、Oracle 和 PeopleSoft 等公司的后端系统之间提供"高速缓存"或数据分级。

数据集成提供了在一个企业主计算机上存储的后端信息的一个镜像。当一个 Internet 客户需要检查一项订单的状态时，这项查询就被转移到数据集成软件。因此，并非总需要访问该企业的主计算机。数据集成软件拥有足够的智能，知道什么时候与主计算机保持同步以便使数据不断更新。为电子商务应用集成 ERP 数据是通过数据分级和直接访问 ERP 数据这两者的结合来完成的，它包括使用一个数据服务器和一些数据高速缓存器。数据集成软件以智能方式将直接实时的和分批的数据存取方法混合起来，以便从一个 ERP 系统中抽取数据。

数据从一个或多个源前进到一个或多个目标表以及信息类型(如 XML)，数据移动的步骤包括确定应该从中抽取数据的源、数据应当进行的转换以及向什么地方发送数据。用户通过一个图形用户接口来指定数据映射和转换。

由用户定义的程序控制每一块数据的移动并确定这种移动之间的内部相关性。例如，如果一个目标表依靠其他目标表的值，则使用一些程序来指定一个数据服务器应当按什么次序来管理这些目标表中的单个数据移动。数据移动可以被设计来以批量方式或实时方式运行，并由管理员来创建和管理，以控制 ERP、电子商务、客户关系管理、供应链管理以及通信应用之间的数据移动。

数据移动使用分布式查询优化、多线程、存储器内数据转换和并行流水线操作来提供很高的数据通过量和可伸缩性。例如，要管理抽取程序并从 SAP 软件中来执行批量数据抽取，可使用优化的 ABAP 代码(SAP 的专有编程语言)，不需要开发和维护定制的 ABAP 代码。

数据集成是企业进一步发展面临的问题。通过数据模型建模和相关应用技术在企业信息集成应用上做了一定的分析。在有效应用模型设计思想开发应用的同时，应重点把

握以下几点。

（1）模型的时效性：包括开发期模型和运行期模型，而运行期模型则显示了模型驱动的核心思想。

（2）模型的进化性：它揭示了模型是否可以根据应用的变化而自我进行改变。

（3）模型的层级性：随着系统的复杂性增加，模型可以由多层级构成。

4．传统数据集成方法的不足

传统数据集成方法存在不足之处。它们不能解决当今 IT 环境的复杂性，也不能覆盖 IT 必须执行的一系列方案的处理。

对于连接数百（或数千）个应用程序的不同单点解决方案，它们仅仅分裂运营数据并将其锁定在部门应用程序中，例如 ERP 和 CRM。以应用程序为中心的数据集成方法没有考虑所有企业数据。例如，它们不能处理计划数据，这些计划数据通常保存在 Excel 电子数据表中，而未保存在部门数据库应用程序中。它们也不能解决驻留在企业外部的有关 BPO 或 SaaS 供应商的数据或与贸易合作伙伴共享的数据。

手动编码数据集成方法也不起作用。手动编码费时费力，并且还容易犯错。由于 IT 机构力求管理更多的数据和更多的数据格式，手动编码通常导致更复杂——而不是更简单。它会增加维护成本并使 IT 效率下降。

在数据质量方面的表现如何？传统数据集成方法无法保证所有数据（客户数据、物料与资产数据以及财务数据）保持完整、一致、准确和最新，而无论数据驻留于何处。

如果继续采用传统方法进行数据集成，即按部门、按应用程序或按数据库，在“孤岛”中进行数据集成，那么有可能需要花费更多时间和金钱来管理复杂情况并“保持业务持续运转”，而不是集中精力来处理新的业务规则。

5．新的数据集成方法的特点

IT 机构需要采用可靠的新方法进行数据集成，这些新方法可以完成如下工作：

- 集成企业内的所有内部预置数据孤岛，包括非结构化数据。
- 集成云计算应用程序和系统中的外部数据。
- 与贸易合作伙伴之间以企业对企业的形式无缝交换数据。
- 确保所有数据的质量。
- 经济高效地管理应用程序生命周期。

数据集成平台是一整套全面的技术，包括访问、发现、清洗、集成并为扩张的企业提供数据。数据集成平台支持各种数据集成项目，例如，数据仓库、数据迁移、测试数据管理、数据存档、数据整合、主数据管理、数据同步、B2B Data Exchange。

6．理想的数据集成平台

数据集成平台必须解决企业间数据碎片的问题，以更快地做出数据驱动型业务决策和更有效地进行业务运作。它必须作为企业技术基础提供服务，提供容易掌控的方法来集成数据。

要满足这些需求，数据集成平台必须具备四个特性：全面、统一、开放和经济。

1）全面

理想的数据集成平台必须具备全面的功能集，使 IT 机构可以根据要求随时随地为企业提供可以信赖的数据。借助一整套可随意支配的数据集成功能，IT 机构的生产效率可以获得数十倍的提升。

2）支持完整的数据集成生命周期

数据集成平台必须支持数据集成生命周期中的所有五个关键步骤：访问、发现、清洗、集成和交付（见图 5.4）。

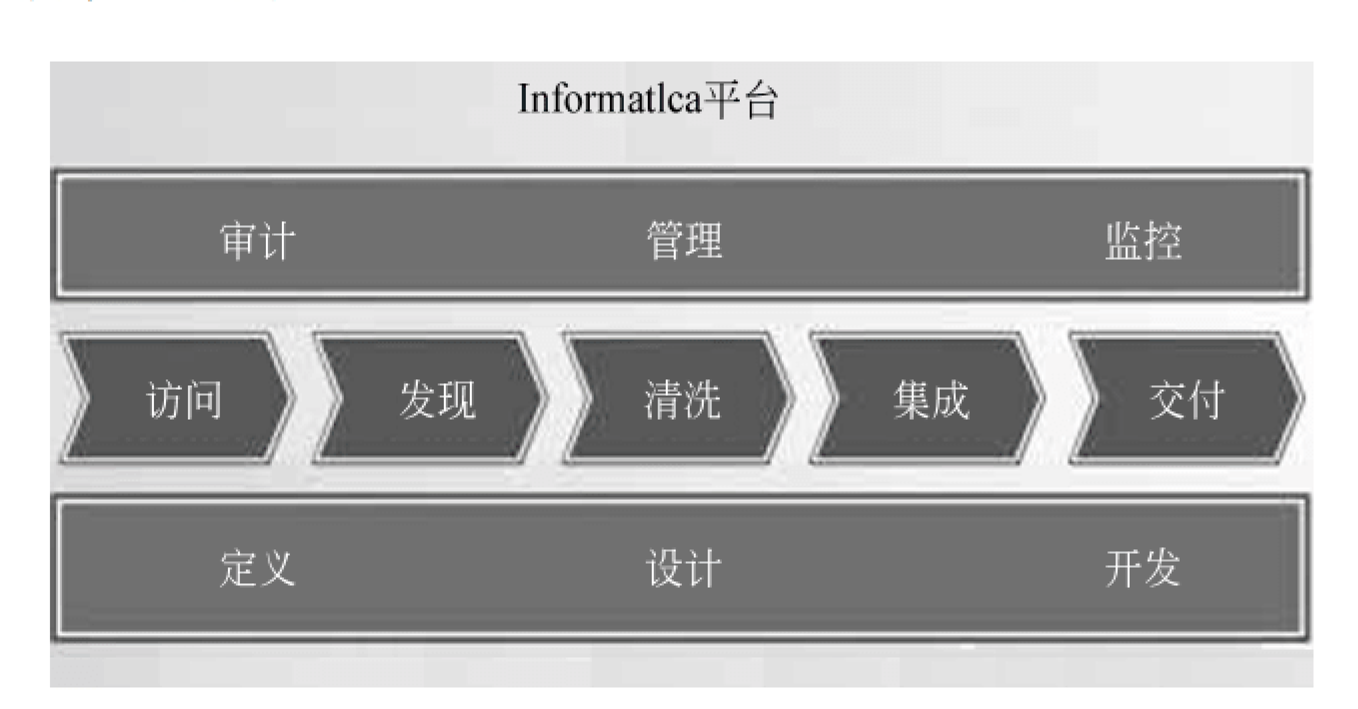


图 5.4 数据集成生命周期

第 1 步：访问。

大多数机构的数据存储在数千个位置，不只限于企业内部，还存放在防火墙外的业务合作伙伴或 SaaS 供应商的“云”中。无论何种来源或结构，所有数据都必须可以接受访问。必须从隐秘的大型主机系统、关系数据库、应用程序、XML、消息甚至从电子数据表之类的文档中提取数据。

第 2 步：发现。

数据源——特别是记录不详尽或来源未知——必须探查才能了解其内容和结构。需要推断数据中隐含的模式和规则。必须标记潜在的数据质量问题。

第 3 步：清洗。

必须清洗数据以确保其质量、准确性和完整性。必须解决错误或疏漏问题。必须强制执行数据标准，并且对值进行验证。必须删除重复的数据条目。

第 4 步：集成。

要跨越多个系统保持一致的数据视图，必须集成并转换数据，以便协调不同系统在定义各种数据元素并使之结构化的方式上存在的差异。例如，对于“客户盈利”，营销系统和财务系统可能具有完全不同的业务定义和数据格式，这些差异必须得到解决。

第 5 步：交付。

必须以适当的格式、在适当的时间将适当的数据交付给所有需要数据的应用程序和用户。交付数据的范围涵盖从支持实时业务运营的单个数据元素或记录到用于趋势分析和企业报告的数百万个记录。必须确保数据的高可用性和交付安全性。

此外，数据集成平台还必须支持如下各部分工作：

（1）审计、管理和监控。

数据管理员和 IT 管理员需要协作进行审计、管理和监控数据。不断地对关键指标

(例如数据质量)进行衡量,随着时间的推移这些指标会得到有目共睹的稳步提高。这是为了跟踪关键数据属性的进度,并标记任何新问题,以便在将数据传回数据集成生命周期之后,可以解决这些问题并不断改进。

(2) 定义、设计和开发。

业务分析师、数据架构师和IT开发人员需要一套功能强大的工具来帮助他们在定义、设计和开发数据集成规则与流程上展开合作。数据集成平台应包括一套常用的集成工具,以确保所有人员一起有效工作。

(3) 数据集成平台必须足够可靠、灵活和可扩展,以处理任何一类型的数据集成项目,其中包括数据仓库、数据迁移、测试数据管理和存档、数据整合、主数据管理、数据同步、B2B Data Exchange。

从单个部门的数据仓库项目到全局数据迁移项目,IT机构可以一次性开展许多类型的数据集成项目。项目团队需要能够从小规模的一个项目类型入手,然后在接下来的项目中重复运用相同的技术和资产——通过共享元数据实现。

(4) 数据集成平台需要能够处理分析数据集成(报告和分析),还要能够处理运营数据集成(与运营执行相关的业务流程)。

(5) 可以在任何周期提供数据。

对于数据集成,存在跨度很广的一系列时间范围和周期要求,这取决于应用程序和使用案例。某些项目要求按月或按周集成数据;而另外一些项目需要按秒提供集成的数据。IT机构需要能够灵活更改周期要求,而不必重新构建整个基础结构。

如图5.5所示,理想的数据集成平台必须在整个周期范围内提供支持、根据应用程序或用户需要随时提供可信任的数据——无论以实时、批量还是变更数据捕获的方式。

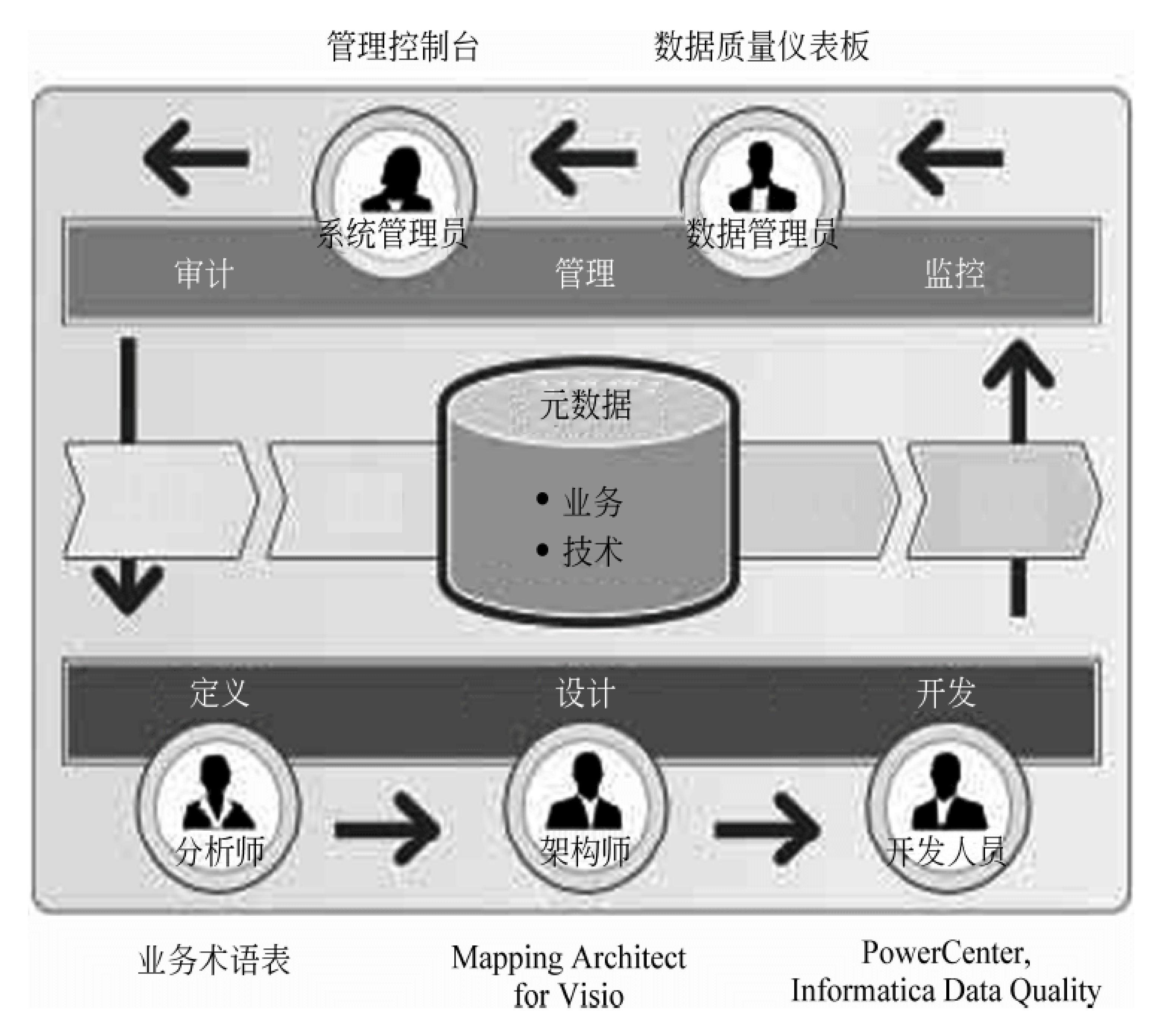


图5.5　基于角色的协作

3）统一

单个的统一数据集成平台可大大简化 IT 团队的工作。当具备扩展型企业（从单一供应商发展成）所需的所有数据集成能力时，通过基于角色的协作、共享元数据和单一的统一运行时引擎，可最大限度地提高工作效率。

（1）基于角色的协作。

数据集成项目包括充当多个角色的 IT 和业务人员。他们都肩负着有待完成、差别很大的任务，可以提供不同的技能。每个角色都需要一套特别为其设计的不同工具。同时，项目团队成员必须精诚合作、共同承担工作和任务，以提高跨团队的工作效率并确保 IT 和业务部门的协调。

如图 5.5 所示，理想的数据集成平台提供角色专用的工具，这些工具专门针对每人的技能和任务而设计。这些角色专用的工具拥有一致的界面。这些工具拥有相同的界面和使用感受，并且相互集成。因此，它们易学易用。通过跨越不同数据集成项目重复使用资产，团队成员能够快速启动运行并保持高效。

（2）共享元数据。

数据集成平台必须提供共享的元数据。平台内的每个工具必须能够访问有关数据存储位置的元数据以及与其关联的业务规则和逻辑。借助共享的元数据，大家可以共同处理同一件事。分析师和开发人员可以处理不同类型的元数据或者用不同方式查看相同的元数据，并仍然保持有效协作。元数据保持一致，并且每个用户均能轻松查看潜在的更改可能带来的影响。

（3）统一的运行时引擎。

数据集成平台的关键是单个的运行时引擎。组成平台的各个单独的产品应全都在简化实施、管理和维护的相同引擎上运行。单个引擎确保可以更为方便地升级多个版本。平台必须为企业级部署而设计，具备可靠的可扩展性、可用性和安全性，这样就可以在该平台上放心开展业务。

4）开放

开放、中立的数据集成平台旨在能够在当前的 IT 环境中兼容一切——硬件、软件、技术标准，以及未来要添加的任何内容。开放的平台能保护企业免于有关供应商瓶颈的风险。

（1）访问任何来源的数据。

大多数机构以数百种不同格式来存储数据：企业应用程序、数据库、平面文件、消息队列、电子数据表和其他文档。数据集成平台必须处理任何数据类型或格式，包括任何来源的结构化和非结构化数据和所有主数据类型，例如客户数据、产品数据和财务数据。

越来越多的数据迁移要跨越公司防火墙和“移入云”。随着更多公司依赖人力资源应用程序和 CRM 应用程序的 SaaS 提供商，云计算变得更为主流。数据集成平台必须能够访问驻留在企业外部的数据。这包括来自多个业务实体的数据和分布在许多不同地理位置和国家/地区的数据。

（2）降低风险。

IT 格局正在改变。这导致不确定性。IT 机构需要采用策略来降低这种变化带来的

风险。你需要一个数据集成平台，它支持从操作系统到数据库的当前所有技术标准。它必须是开放式的，确保能够与现有或将来可能配置的一切内容兼容。这包括在企业与“云”中或合作伙伴的全部各种应用程序和数据源。

5）经济

经济的数据集成平台能够带来尽可能低的总拥有成本（TCO）和最快、最高的投资回报（ROI）。在当前严峻的经济环境下，现在和将来的每笔技术投资都要接受严格审查，评估其帮助 IT 机构和业务的能力，因此这些因素目前显得特别重要，主要涉及的因素有降低成本、更为高效地运营、快速产生价值、更低的总拥有成本。

经济的数据集成平台能够获得更快的投资回报。

在数据集成平台中获得快速的投资回报取决于能否迅速行动并投入使用。从而需要增加 IT 资源。

5.3.2　数据集成方案

继系统集成、应用集成、业务集成之后，数据集成（Data Integration，DI）已渐被各大企业纷纷触及。目前国内大多数企业还仅停留在服务于单个系统的多对一架构数据集成应用，这种架构常见于数据仓库系统领域，服务于企业的商务智能。早期那些数据集成大家大都是从 ETL 启蒙开始的，当时 ETL 自然也就成了数据集成的代名词，只是忽如一夜春风来，各厂商相继推出 DI 新概念后，我们不得不再次接受新一轮的 DI 洗脑，首推的有 SAS DI、Business Objects DI、Informatica DI、Oracle DI（ODI）等厂商。

数据集成主要是指基于企业分散的信息系统的业务数据进行再集中、再统一管理的过程，是一个渐进的过程，只要有新的、不同的数据产生，就不断有数据集成的步骤执行。企业经历了几年的信息化发展，凌乱、重复、歧义的数据接踵而至，数据集成的空间与需求日渐迫切，企业需要一个主数据管理（Master Data Manager）系统来统一企业的产品信息、客户信息；企业需要一个数据仓库（Data Warehouse）系统来提高领导层的决策意识，加快市场战略调整行动；企业需要一个数据中心（Data Center）系统来集中交换、分发、调度、管理企业基础数据。

数据集成的必要性、迫切性不言而喻，不断被推至企业信息化战略规划的首要位置。要实现企业数据集成的应用，不仅要考虑企业急需集成的数据范围，还要从长远发展考虑数据集成的架构、能力和技术等方面内容。从数据集成应用的系统部署、业务范围、实施成熟性看主要可分为三种架构：单个系统数据集成架构、企业统一数据集成架构、机构之间数据集成架构。

1. 单个系统数据集成架构

单个系统数据集成架构是国内目前应用最广的架构，主要是以数据仓库系统为代表提供服务而兴建的数据集成平台，面向企业内部如 ERP、财务、OA 等多各业务操作系统，集成企业所有基础明细数据，转换成统一标准，按星型结构存储，面向市场经营分析、客户行为分析等多个特有主题进行商务智能体现。这种单个系统数据集成应用架构的主要特点是多对一的架构、复杂的转换条件、TB 级的数据量处理与加载，数据存储结构特殊，星

型结构、多维立方体并存,数据加载层级清晰。单个系统数据集成架构见图 5.6 所示。

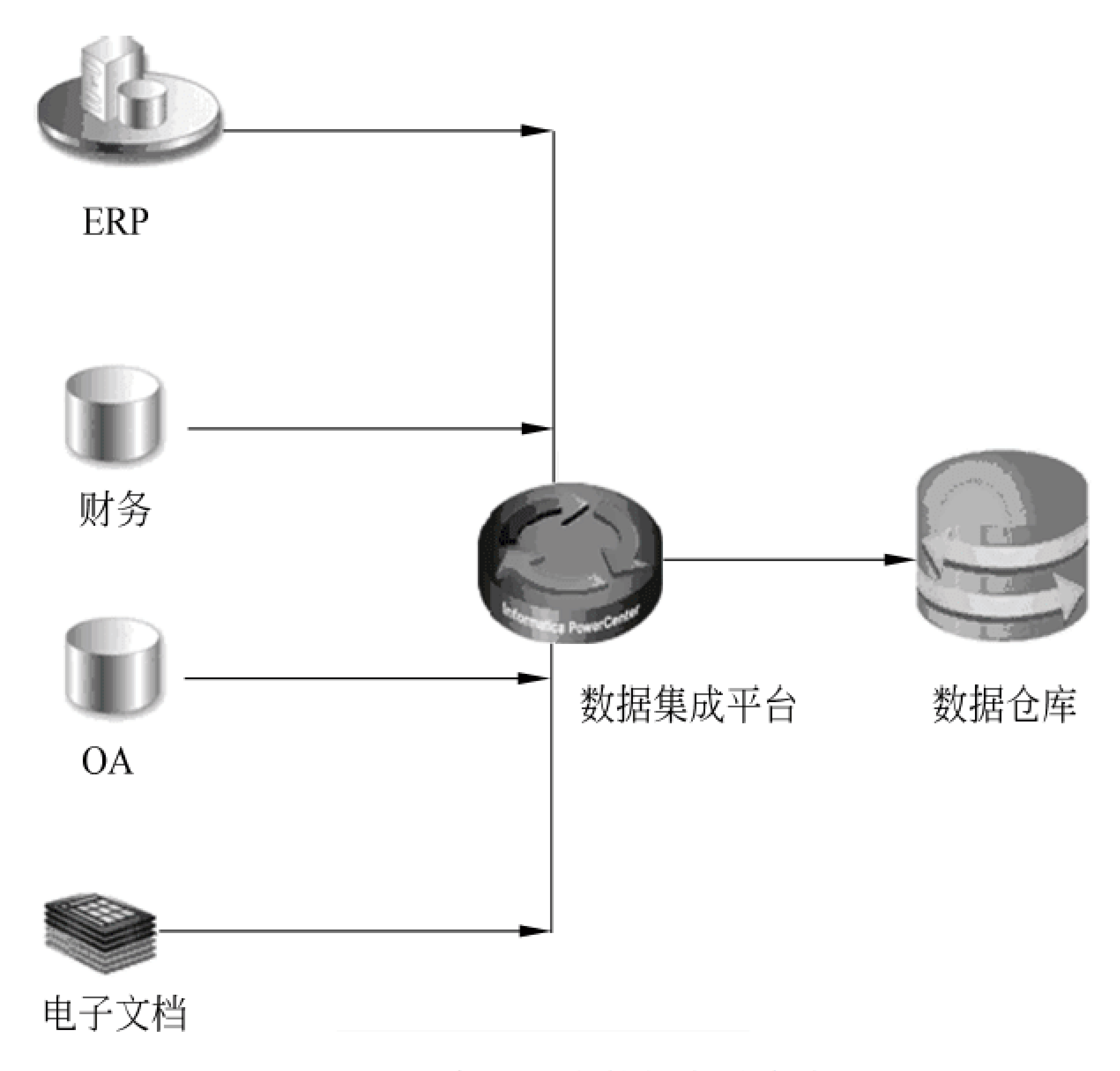


图 5.6 单个系统数据集成架构

2. 企业统一数据集成架构

组织结构较复杂的大型企业、政府机构尤为偏爱这种数据集成的架构,因此类单位具有业务结构相对独立、数据权力尤为敏感、数据接口复杂繁多等特征,更需要多个部门一起协商来建立一个统一的数据中心平台,以满足部门之间频繁的数据交换的需求。如金融机构,电信企业,公安、税务等政府机构,业务独立、层级管理的组织结构决定了内部数据交互的复杂性。概括来说,此类应用属于多对多的架构、数据交换频繁、要有独立的数据交换存储池、数据接口与数据类型繁多等特点。

对于企业管理性、决策性较强的信息系统,如主数据管理系统、财务会计管理系统、数据仓库系统等数据可直接来源于数据中心,摆脱了没有企业数据中心前的一对多交叉的困扰,避免了业务系统对应多种管理系统时需要数据重复传送,如 CRM 系统中新增一条客户信息数据后,直接发送到企业数据中心,由企业数据中心面向风险管理系统、数据仓库系统、主数据管理系统进行分发即可。

企业统一数据集成架构见图 5.7 所示。

3. 机构之间数据集成架构

这种架构多是应用于跨企业、跨机构、多个单位围绕某项或几项业务进行的业务活动,或由一个第三方机构来进行协调这些企业、机构之间的数据交换、制定统一数据标准,从而形成一个多机构之间的数据集成平台。如中国银联与各商业银行之间的应用案例、各市政府信息中心与市政府各机关单位之间的应用案例、外贸 EDI(海关、检验检疫局、外汇局、银行、保险、运输等)、BTOB 电子商务平台等。这类应用属于跨多企业、单位多对多的架构,具有数据网络复杂、数据安全性要求高、数据交换实时性强等特点。

尤其这类架构颇具一些特点值得进一步去剖析。因数据集成平台是架于多企业、单

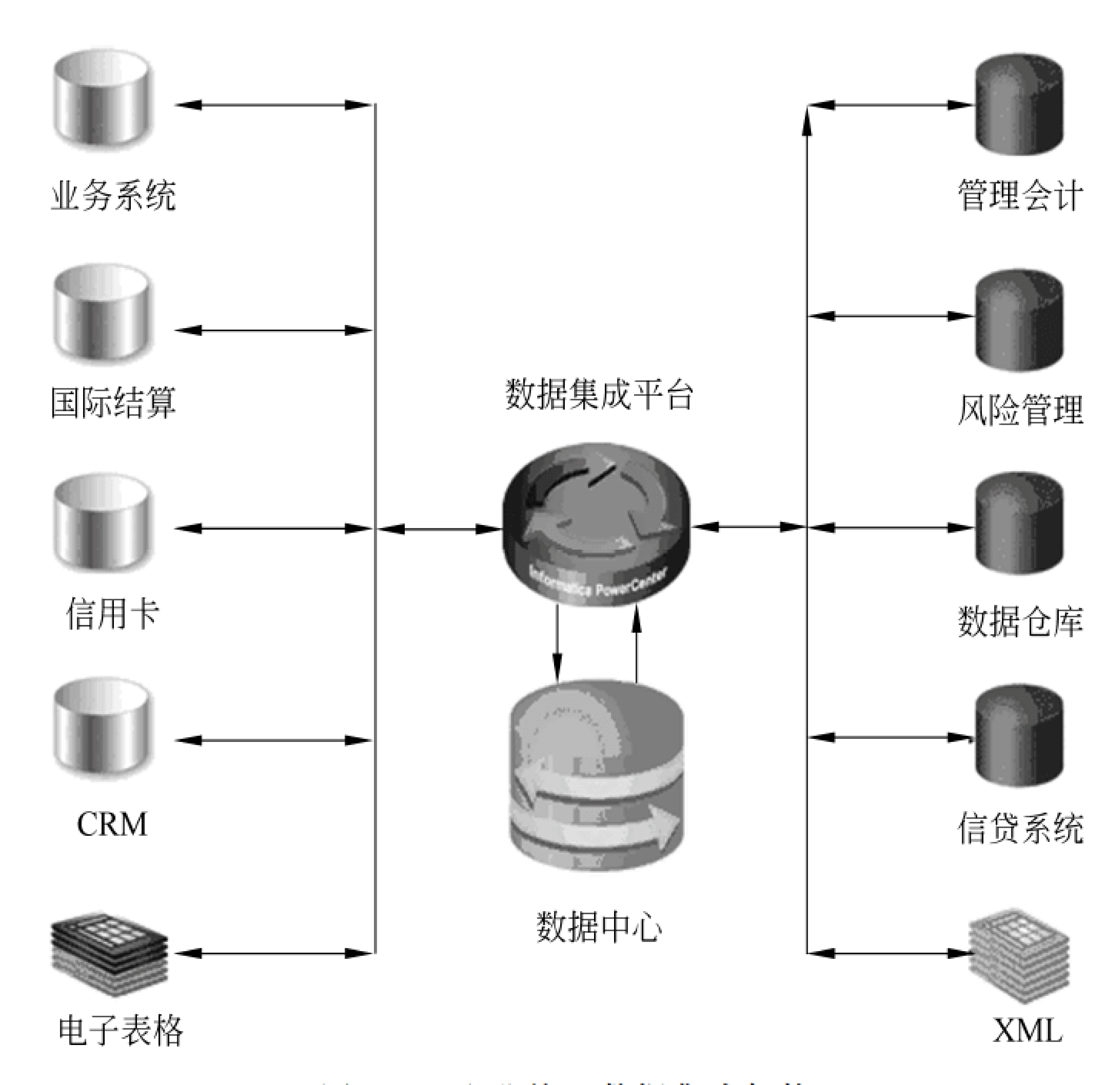


图 5.7　企业统一数据集成架构

位之间，数据的安全性、独立性决定了各企业、单位不得不考虑前置机的部署形式，各企业、单位在业务系统与数据集成平台之间增加一台前置机，则更有利于自有系统数据的独立与安全，也更利于数据平台对数据的获取、分发、交换的统一要求。另外，数据集成平台也要具有更多的技术功能来满足众多单位的众多数据接口、多种数据类型、不一致的数据标准、数据交换的实时性、对数据的抽取与推送(Pull AND Push)等业务需求。如数据集成平台需具有数据连通、ETL、数据实时、数据清洗、数据质量、企业服务总线(Enterprise Service Bus，ESB)、面向服务的体系结构(Service-Oriented Architecture，SOA)等一些技术与特点。

机构之间数据集成架构如图 5.8 所示。

以上三种数据集成架构，一种是对应于某一个应用系统的多对一架构，一种是完成企业内部众多系统之间数据交换的多对多架构，一种是为多个跨企业、单位机构实现某一项或几项业务活动而建立的多对多架构，数据集成的应用差不多都是基于这三种架构，每种架构可能会对应多种数据集成的应用。国内企业常见的数据集成应用有数据仓库、数据同步、数据交换，随着企业并购、新旧系统升级、分布系统向数据大集中看齐、电子商务的发展、多个企业单位协同作业等等众多业务需求的诞生，数据集成的应用开始纷繁异景起来。

5.3.3　企业数据集成应用形式

目前大部分数据集成软件厂商都是围绕数据仓库(Data Warehousing)、数据迁移(Data Migration)、数据合并(Data Consolidation)、数据同步(Data Synchronization)、数据交换(Data Hubs 或者叫主数据管理：Master Data Management)这 5 种常见的企业应

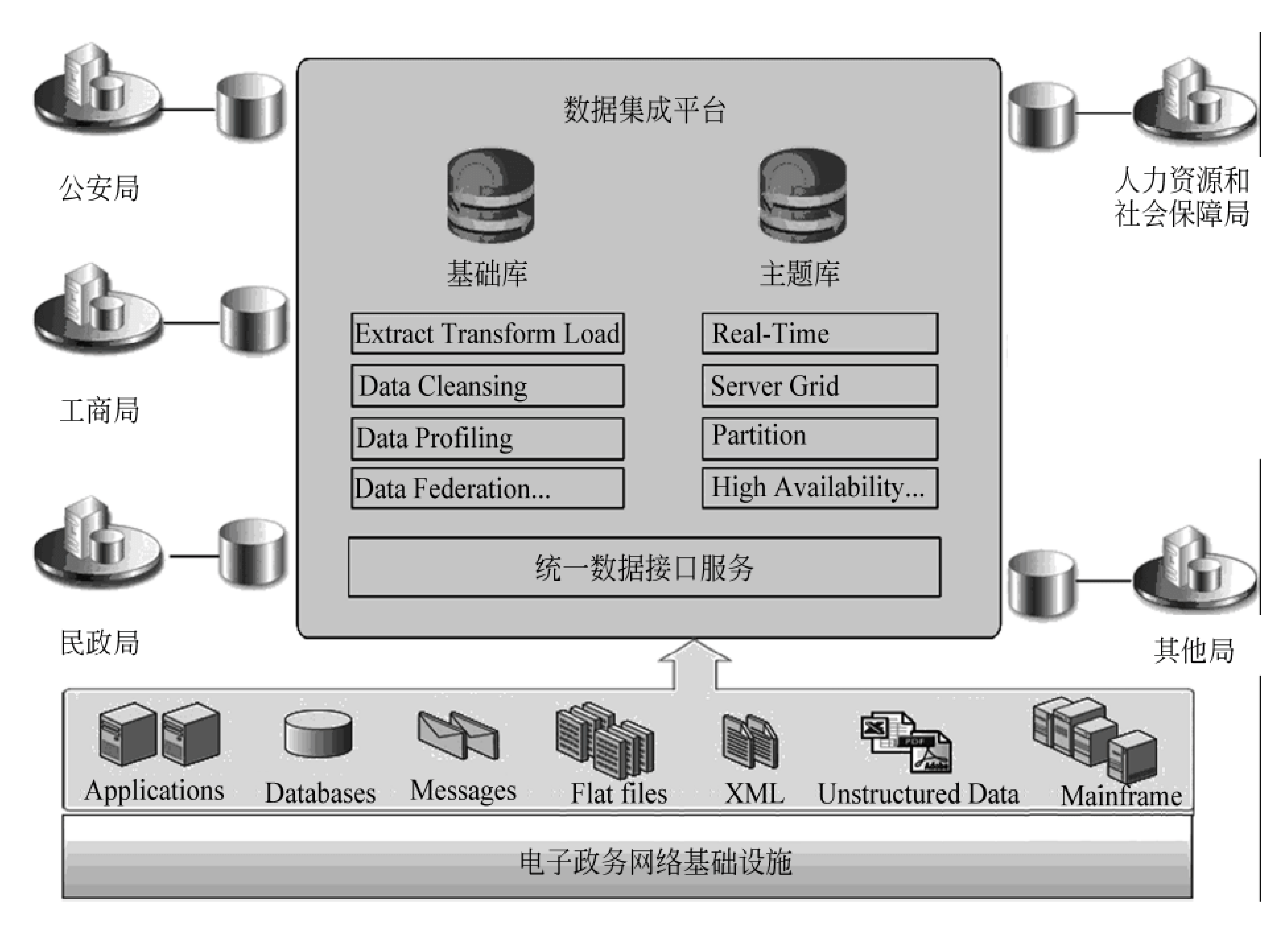


图 5.8　机构之间数据集成架构

用形式来发展各自的产品技术。

1. 数据仓库(Data Warehousing)应用

数据仓库中的数据集成应用主要是围绕 ETL 的功能来实现,一般来说其主要功能是将多个业务系统不同种数据类型的数据抽取到数据仓库的 ODS(Operational Data Store)层,经过转换,加载存储到星型结构的 DW(Data Warehouse)层,为满足不同主题的展现应用,再向关系型数据库或多维数据库进一步汇总加载,其 ETL 功能可由手工编程或专业工具软件这两种类型来实现。数据仓库应用如图 5.9 所示。

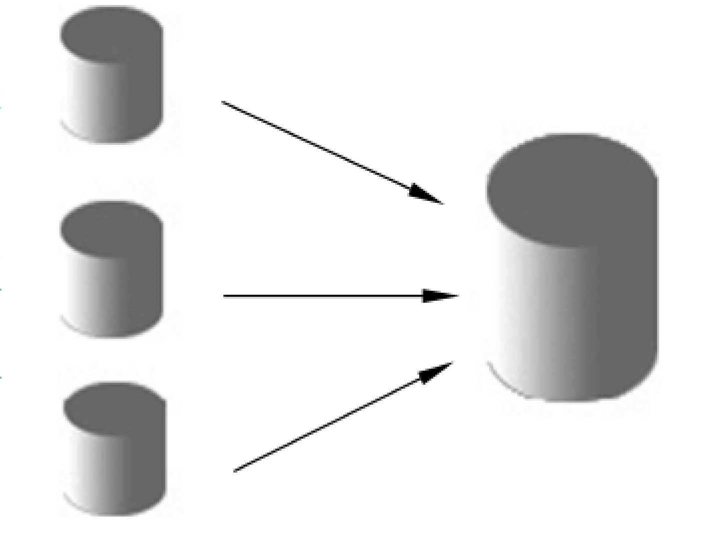


图 5.9　数据仓库应用

第一种类型:由手工编程到专项 ETL 工具的应用,这种应用类型是成熟的数据集成软件工具的雏形,是为快速达成项目功能需求为主,满足复杂的业务处理的需要,以 ETL 为核心应用,开发技术也发挥得淋漓尽致,PB、Java、SQL、存储过程、C/C++ 都可能会悉数登场,多一种系统的数据集成就可能会有多于一倍的开发工作量,使数据集成平台更趋于复杂、脆弱。另外,如电信、金融、税务、公安等行业的众多系统集成商针对各自的业务系统也开发有专项的数据集成工具,只是有一定的局限性,拘泥于某一种应用或某一特定的系统环境。

第二种类型:众多成熟的数据集成软件工具的应用为这一代表,如 Informatica PowerCenter、IBM Datastage、Oracle ODI、Microsoft SISS 等,集各种数据接口、ETL、数据质量、实时、数据联邦、分区并行、网格、HA 等技术于一身,历练世界众多客户需求多

时，具有更宽广的应用、可扩展性强、安全稳定等一些特点。

2. 数据迁移(Data Migration)应用

这种应用比较容易理解，对于新旧系统升级、数据大集中时的数据作迁移，使数据更能顺应新系统的结构变化而平稳迁移。数据迁移应用如图5.10所示。

3. 数据合并(Data Consolidation)应用

在企业并购中很容易产生数据合并的应用，如两个企业的HR系统的合并、财务系统的合并、其他业务系统的合并，当系统需要合并必然产生数据的合并，因此对企业数据进行统一标准化、规范化、数据的补缺、数据的一致性都将导致数据合并。数据合并应用见图5.11所示。

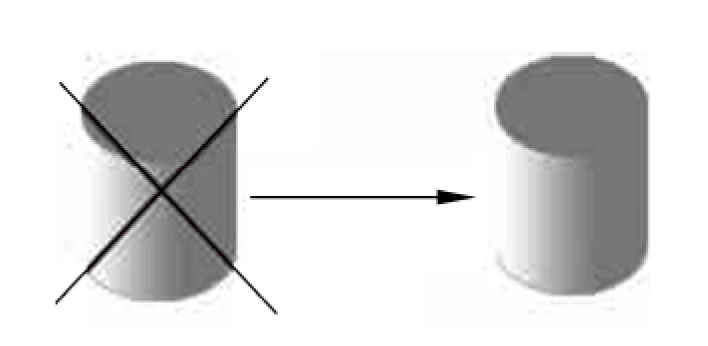


图5.10　数据迁移(Data Migration)应用

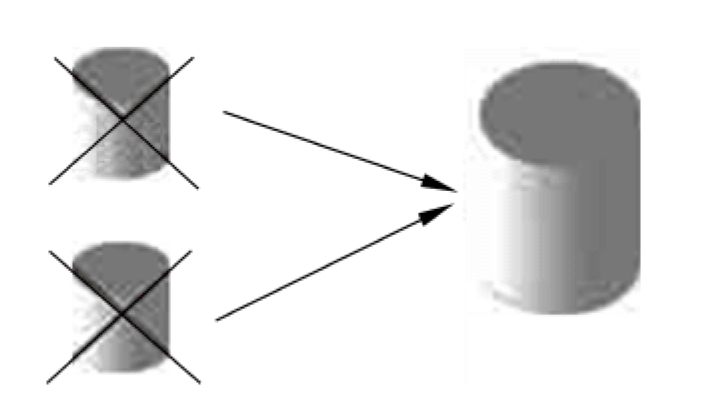


图5.11　数据合并应用

4. 数据同步(Data Synchronization)应用

当企业一个系统的业务活动会影响其他多个系统的进程时，对数据的实时性、准确性就显得尤为重要。如航空公司与航空机场之间的数据同步应用、证券交易所与证券公司之间的股票信息同步、金融业的汇率信息同步等等，影响数据同步的实时性与可靠性的因素会有网络的连通性、传输效率、数据接口、数据格式等，这些诸多因素都属于数据集成中的数据同步要解决的问题。数据同步应用见图5.12所示。

5. 数据交换(Data Hubs)应用

或者叫主数据管理(Master Data Management)应用，这种数据集成的应用越来越受到企业的重视。一般构成企业主要的基础数据分别是客户数据、产品数据、员工信息数据、供应商数据，要从企业多个系统中快速、可靠地建立唯一、完整的企业主数据视图，这就是主数据管理。要实现企业主数据管理应用的数据集成平台，必须具备有良好的数据连通性、良好的数据质量探查与分析、良好的数据转换能力等特点。数据交换应用如图5.13所示。

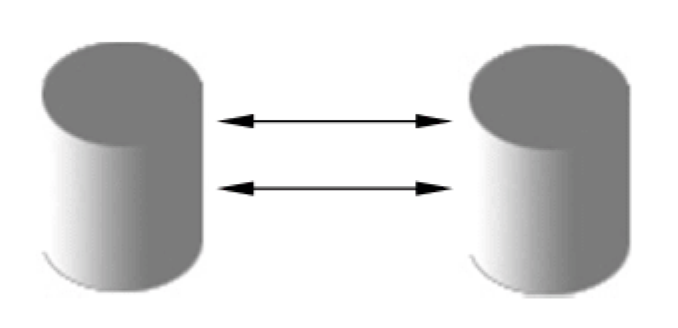


图5.12　数据同步应用

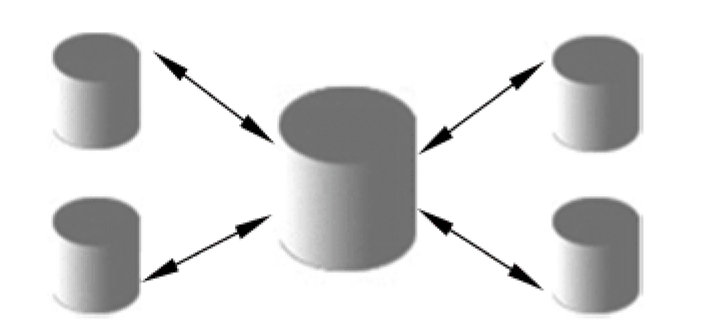


图5.13　数据交换应用

上述提到跨多个企业、单位机构的架构就是一个典型的主数据管理应用，如公安局、工商局、税务局、人事局、劳动社保局等这些众多政府机构主要是围绕两个基本主体进行

各项事务活动：一个主体是个人，另外一个主体是企业单位。众多政府机构对这两个主体的信息数据要求重点不同、数据处理顺序有先后，数据变更有各异，数据交换复杂、频繁，而最理想的境界是这两个主体数据能做到最大程度的同步，这就是主数据管理的思想。

以上五种数据集成应用解决方案在国内最常见的是数据仓库的应用，最复杂的应用应该是数据交换了，不管是简单还是复杂的应用都以 ETL 技术为基础，ETL 技术成为了数据集成的核心技术，伴随 ETL 技术的还有数据连通、数据质量、数据清洗、数据联邦、Real-Time、数据探查等技术，为了提高数据集成的安全性、高效性、可扩展能力，还有 SOA、HA、GRID 等相关技术作为支撑。

1）ETL（Extract、Transform、Load）

数据集成视数据抽取、转换和加载为最基础、最核心的三项技术，这三个执行步骤可根据系统环境特点调整顺序，典型的应用有 ELT 的顺序。如源与目标为同种数据库或共用一个数据库时，可将数据从源直接抽取到目标然后再进行转换，效率会大有提高，专注此类特点的产品以 Oracle 的 ODI 为代表。

2）数据连通（Data Connective）

良好的数据连通性是数据集成的能力体现，一般通用的关系型数据库、ODBC、XML 等数据连通类型为常见类型，还有一些就是大中型企业常用的 ERP、CRM、BPM、OA 等应用软件为封闭式的系统，如 SAP、Seibel、Lotus 等系统的连通，因此良好的数据集成平台需要提供来自更多企业的数据连通接口，抽取源与装载目标的范围也就更广阔。

3）数据质量（Data Quality）

数据质量越来越被企业重视，数据质量的技术范围也越发宽广，开始慢慢被剥离出数据集成的范畴。企业不能根据标准不统一、歧义、不正确的数据快速做出决策，只有站在高质量的数据基础之上做出的决策才不会发生方向偏倚。通常实现企业数据质量管理会包括源数据的探查、数据质量的评估、数据集成、数据的完整和数据的监控这五个步骤。数据的完整一般是指根据现有基础数据作其他数据项的扩展和丰富，如根据客户的联系方式来丰富客户的所属地区数据项、根据客户身份证号码来丰富客户的所属地区、年龄、性别等信息。

4）数据实时（Real-Time）

对于实时数据仓库系统、数据同步等应用都会用到数据实时技术，一个系统的数据发生变化后，能即刻将变化的动作同步到另一个系统这就是数据实时技术的主旨。关系型数据库、AS400、MQ Series、ADABAS 等系统都有自身的实时数据策略，如 Oracle 数据库的实时技术可以通过 Trigger 或 Log Miner 分析归档日志方式来实现。

诸如以上 ETL、数据连通、数据质量、数据实时等技术，还有数据联邦、数据清洗、HA、Grid、Partition、SOA 技术，这些都是保证数据集成平台的可扩展性、安全性、高效性、简便性的通用技术。

5.3.4 企业整体解决方案

常见的整体解决方案包括有企业数据集成业务咨询、企业数据集成平台产品、各厂商

数据集成底层软件共三大块。图5.14给出了神州数码数据集成解决方案示意。

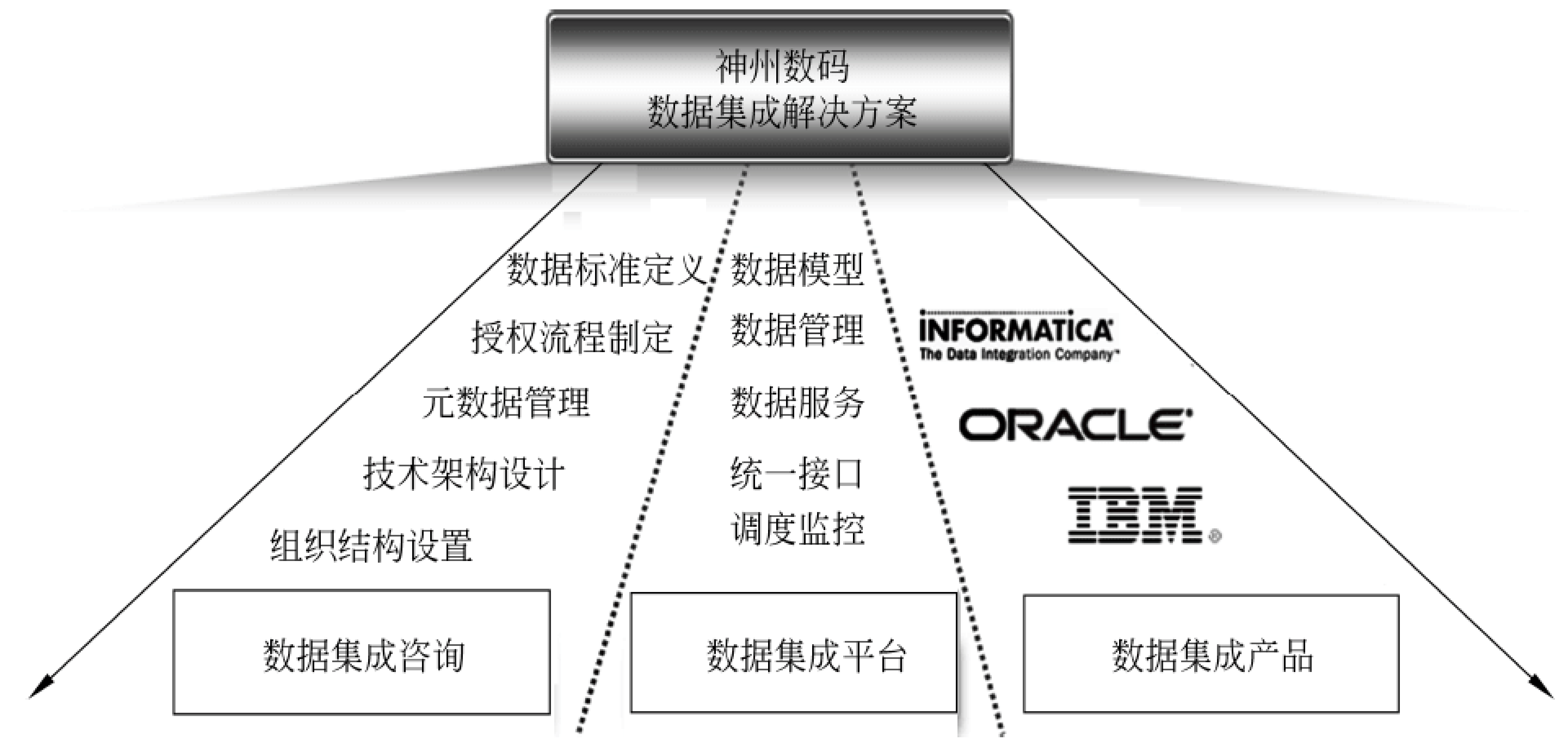


图5.14 神州数码数据集成解决方案

1. 数据集成咨询

业务咨询具体指对企业各个层次的数据对象进行调研，给出企业数据管理现状分析报告，为企业的数据管理进行数据标准定义，根据企业特点提出更优的核心数据管理机制建议，设计适合企业长远发展的数据管理机构体系和工作管理流程，并对组织结构进行岗位职能设置。

2. 数据集成平台

数据集成平台是企业数据管理部门的工作手段，须依赖于一套严谨的数据管理规范。数据集成平台是以企业数据统一存储模型作为依托，提供完备的数据存取、清洗、转换等处理功能，为企业各业务部门提供准确、单一的数据服务，并对数据服务各环节进行审批、监控、分析和管理。

3. 数据集成产品

提供基于客户需求的，以应用软件为核心的IT服务，包括IBM、Oracle、Informatica等厂商的数据集成软件产品。

5.4 机器学习

机器学习这个词是让人疑惑的，首先它是英文名称Machine Learning（简称ML）的直译，在计算界Machine一般指计算机。这个名字使用了拟人的手法，说明了这门技术是让机器"学习"的技术。但是计算机是"死"的，怎么可能像人类一样"学习"呢？

传统上如果我们想让计算机工作，我们给它一串指令，然后它遵照这个指令一步步执行下去。有因有果，非常明确。但这样的方式在机器学习中行不通。机器学习根本不接受你输入的指令，相反，它接受你输入的数据！也就是说，机器学习是一种让计算机利用数据而不是指令来进行各种工作的方法。这听起来非常不可思议，但从结果看来却是非

常可行的。"统计"思想将在你学习"机器学习"相关理念时无时无刻不伴随在旁边，相关而不是因果的概念将是支撑机器学习能够工作的核心概念。你会颠覆对你以前所有程序中建立的因果无处不在的根本理念。

5.4.1 机器学习的定义和例子

从广义上来说，机器学习是一种能够赋予机器学习的能力以此让它完成直接编程无法完成的功能的方法。但从实践的意义上来说，机器学习是一种通过利用数据，训练出模型，然后使用模型预测的一种方法。

让我们具体看一个例子。

拿房子来说，现在我手里有一栋房子需要售卖，我应该给它标上多大的价格？房子的面积是 100 平方米，价格是 100 万元、120 万元，还是 140 万元？

很显然，我希望获得房价与面积的某种规律。那么我该如何获得这个规律？用报纸上的房价平均数据么？还是参考别人面积相似的？无论哪种，似乎都并不是太靠谱。

我现在希望获得一个合理的，并且能够最大程度地反映面积与房价关系的规律。于是我调查了周边一些类似的房子，获得了一组数据。这组数据中包含了大大小小房子的面积与价格，如果我能从这组数据中找出面积与价格的规律，那么我就可以得出房子的价格，如图 5.15 所示。

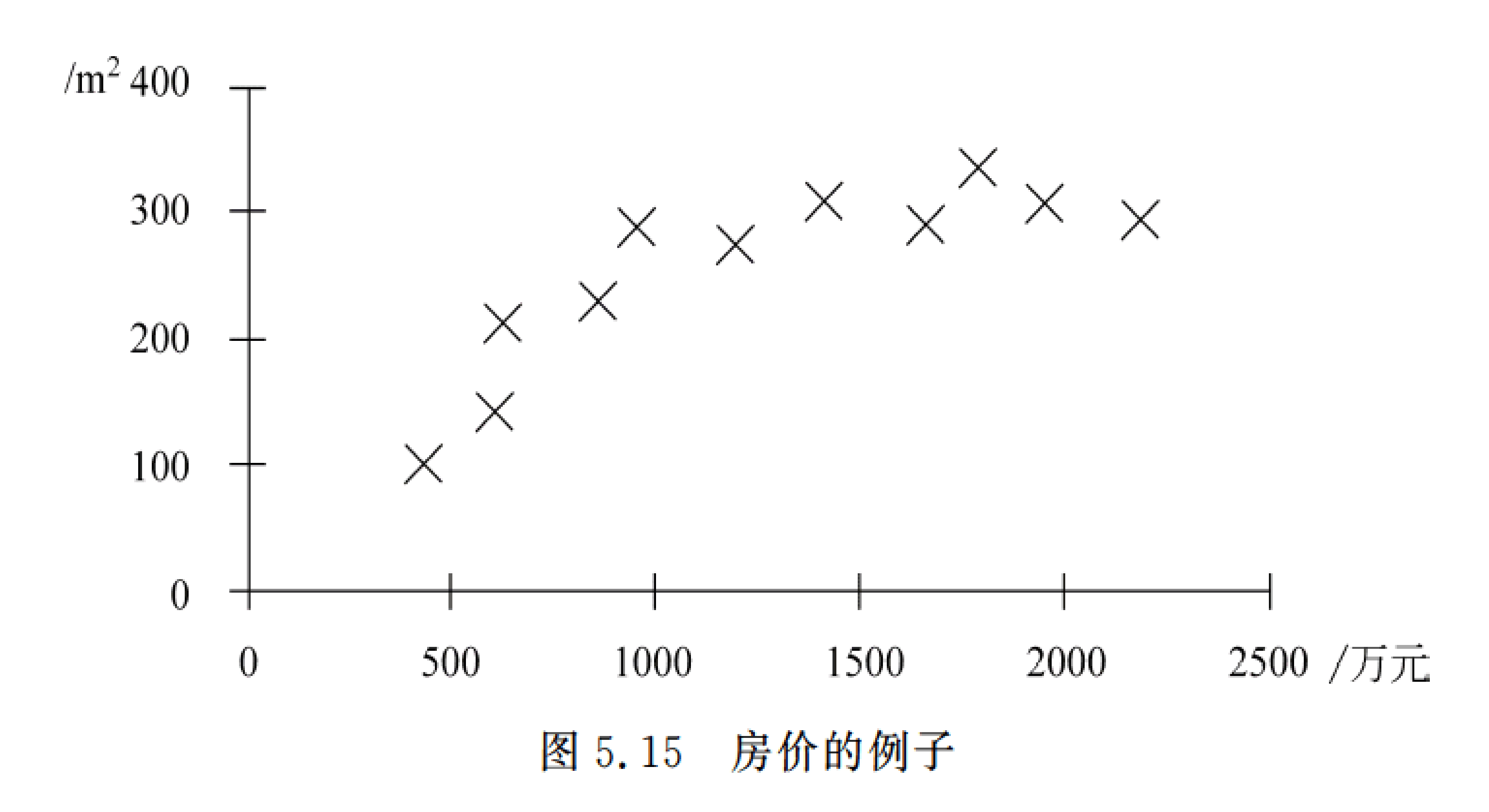


图 5.15 房价的例子

对规律的寻找很简单，拟合出一条直线，让它"穿过"所有的点，并且与各个点的距离尽可能小。

通过这条直线，我获得了一个能够最佳反映房价与面积关系的规律。这条直线同时也是一个下式所表明的函数：

$$房价 = 面积 \times a + b$$

上述 a、b 都是直线的参数。获得这些参数以后，就可以计算出房子的价格。

假设 $a=0.75$，$b=50$，则房价$=100\times0.75+50=125$ 万。这个结果与我前面所列的 100 万、120 万、140 万都不一样。由于这条直线综合考虑了大部分的情况，因此从"统计"意义上来说，这是一个最合理的预测。

在求解过程中透露出了两个信息：

(1) 房价模型是根据拟合的函数类型决定的。

如果是直线，那么拟合出的就是直线方程。如果是其他类型的线，例如抛物线，那么拟合出的就是抛物线方程。机器学习有众多算法，一些强力算法可以拟合出复杂的非线性模型，用来反映一些直线所不能表达的情况。

(2) 数据越多，模型就能够考虑到越多的情况，由此对于新情况的预测效果可能就越好。

这是机器学习界“数据为王”思想的一个体现。一般来说(不是绝对)，数据越多，最后机器学习生成的模型预测的效果越好。

通过拟合直线的过程，可以对机器学习过程做一个完整的回顾。首先，需要在计算机中存储历史的数据。接着，将这些数据通过机器学习算法进行处理，这个过程在机器学习中叫做“训练”，处理的结果可以被用来对新的数据进行预测，这个结果一般称之为“模型”。对新数据的预测过程在机器学习中叫做“预测”。“训练”与“预测”是机器学习的两个过程，“模型”则是过程的中间输出结果，“训练”产生“模型”，“模型”指导“预测”。

让我们把机器学习的过程与人类对历史经验归纳的过程做个比对，如图 5.16 所示。

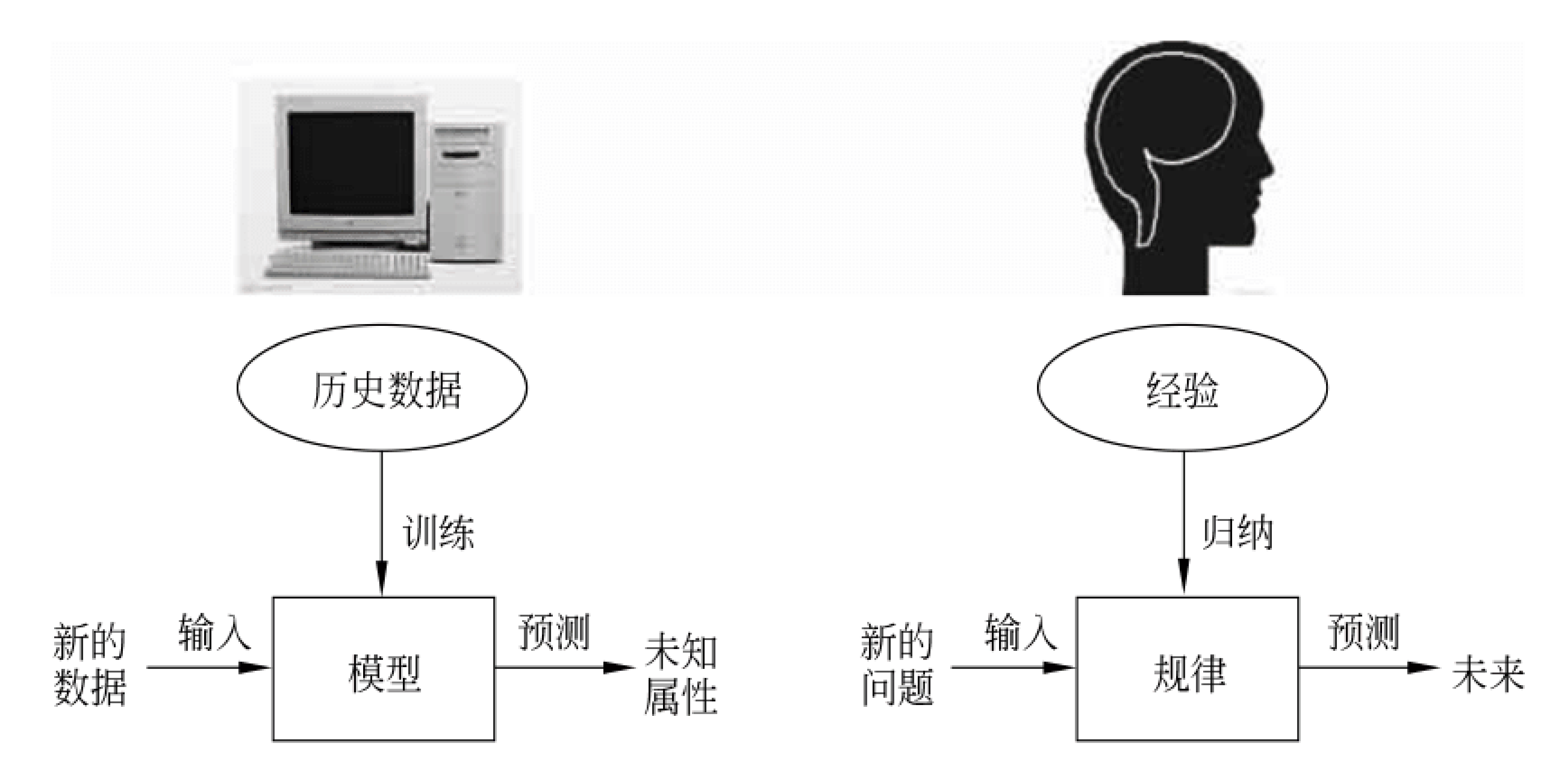


图 5.16　机器学习与人类思考的类比

人类在成长、生活过程中积累了很多的历史与经验。人类定期地对这些经验进行“归纳”，获得了生活的“规律”。当人类遇到未知的问题或者需要对未来进行“推测”的时候，人类使用这些“规律”，对未知问题与未来进行“推测”，从而指导自己的生活和工作。

机器学习中的“训练”与“预测”过程可以对应到人类的“归纳”和“推测”过程。通过这样的对应，我们可以发现，机器学习的思想并不复杂，仅仅是对人类在生活中学习成长的一个模拟。由于机器学习不是基于编程形成的结果，因此它的处理过程不是因果的逻辑，而是通过归纳思想得出的相关性结论。

这也可以联想到人类为什么要学习历史，历史实际上是人类过往经验的总结。有句话说得很好——“历史往往不一样，但历史总是惊人的相似”。通过学习历史，我们从历史中归纳出人生与国家的规律，从而指导我们的下一步工作，这是具有极大价值的。当前一些人忽视了历史的本来价值，而是把其作为一种宣扬功绩的手段，这其实是对历史真实价值的一种误用。

5.4.2 机器学习的范围

上面虽然说明了机器学习是什么,但是并没有给出机器学习的范围。

其实,机器学习跟模式识别、统计学习、数据挖掘、计算机视觉、语音识别、自然语言处理等领域有着很紧密的联系。

从范围上来说,机器学习跟模式识别、统计学习、数据挖掘是类似的,同时,机器学习与其他领域的处理技术的结合,形成了计算机视觉、语音识别、自然语言处理等交叉学科。因此,一般说数据挖掘时,可以等同于说机器学习。同时,我们平常所说的机器学习应用,应该是通用的,不仅仅局限于结构化数据,还有图像、音频等应用。

本节对机器学习这些相关领域的介绍有助于我们理清机器学习的应用场景与研究范围,更好地理解后面的算法与应用层次。

图 5.17 是机器学习所涉及的一些相关范围的学科与研究领域。



图 5.17 机器学习与相关学科

1. 模式识别

模式识别=机器学习。两者的主要区别在于前者是从工业界发展起来的概念,后者则主要源自计算机学科。在著名的 *Pattern Recognition And Machine Learning* 这本书中,Christopher M. Bishop 在开头是这样说的:"模式识别源自工业界,而机器学习来自于计算机学科。不过,它们中的活动可以被视为同一个领域的两个方面,同时在过去的 10 年间,它们都有了长足的发展。"

2. 数据挖掘

数据挖掘=机器学习+数据库。这几年数据挖掘的概念实在是耳熟能详。但凡说到数据挖掘都会吹嘘数据挖掘如何如何,例如从数据中挖出金子,以及将废弃的数据转化为价值等等。但是,我尽管可能会挖出金子,但我也可能挖的是"石头"啊。这个说法的意思是,数据挖掘仅仅是一种思考方式,告诉我们应该尝试从数据中挖掘出知识,但不是每个数据都能挖掘出金子的,所以不要神话它。一个系统绝对不会因为上了一个数据挖掘模块就变得无所不能,恰恰相反,一个拥有数据挖掘思维的人员才是关键,而且他还必须对数据有深刻的认识,这样才可能从数据中导出模式指引业务的改善。大部分数据挖掘中

的算法是机器学习的算法在数据库中的优化。

3. 统计学习

统计学习约等于机器学习。统计学习是个与机器学习高度重叠的学科。因为机器学习中的大多数方法来自统计学，甚至可以认为，统计学的发展促进机器学习的繁荣昌盛。例如著名的支持向量机算法，就是源自统计学科。但是在某种程度上两者是有分别的，这个分别在于：统计学习者重点关注的是统计模型的发展与优化，偏数学；而机器学习者更关注的是能够解决问题，偏实践，因此机器学习研究者会重点研究学习算法在计算机上执行的效率与准确性的提升。

4. 计算机视觉

计算机视觉＝图像处理＋机器学习。图像处理技术用于将图像处理为适合进入机器学习模型中的输入，机器学习则负责从图像中识别出相关的模式。计算机视觉相关的应用非常多，例如百度识图、手写字符识别、车牌识别等等应用。这个领域是应用前景非常火热的，同时也是研究的热门方向。随着机器学习的新领域深度学习的发展，大大促进了计算机图像识别的效果，因此未来计算机视觉界的发展前景不可估量。

5. 语音识别

语音识别＝语音处理＋机器学习。语音识别就是音频处理技术与机器学习的结合。语音识别技术一般不会单独使用，一般会结合自然语言处理的相关技术。目前的相关应用有苹果的语音助手 siri 等。

6. 自然语言处理

自然语言处理＝文本处理＋机器学习。自然语言处理技术主要是让机器理解人类的语言的一门领域。在自然语言处理技术中，大量使用了编译原理相关的技术，例如词法分析、语法分析等等，除此之外，在理解这个层面，则使用了语义理解、机器学习等技术。作为唯一由人类自身创造的符号，自然语言处理一直是机器学习界不断研究的方向。按照百度机器学习专家余凯的说法“听与看，说白了就是阿猫和阿狗都会的，而只有语言才是人类独有的”。如何利用机器学习技术进行自然语言的深度理解，一直是工业和学术界关注的焦点。

可以看出机器学习在众多领域的外延和应用。机器学习技术的发展促使了很多智能领域的进步，改善着人们的生活。

5.4.3　机器学习的方法

通过上节的介绍，我们了解了机器学习的大致范围，那么机器学习里面究竟有多少经典的算法呢？本节将简要介绍一下机器学习中的经典代表方法。这部分介绍的重点是这些方法内涵的思想，数学与实践细节不会在这里讨论。

1. 回归算法

在大部分机器学习课程中，回归算法都是介绍的第一个算法。原因有两个：第一，回

归算法比较简单，介绍它可以让人平滑地从统计学迁移到机器学习中；第二，回归算法是后面若干强大算法的基石，如果不理解回归算法，无法学习那些强大的算法。回归算法有两个重要的子类：即线性回归和逻辑回归。

一个线性回归的例子就是我们前面说过的房价求解问题。如何拟合出一条直线最佳匹配我所有的数据？一般使用“最小二乘法”来求解。“最小二乘法”的思想是这样的，假设我们拟合出的直线代表数据的真实值，而观测到的数据代表拥有误差的值。为了尽可能减小误差的影响，需要求解一条直线使所有误差的平方和最小。最小二乘法将最优问题转化为求函数极值问题。函数极值在数学上我们一般会采用求导数为 0 的方法。但这种做法并不适合计算机，可能求解不出来，也可能计算量太大。

计算机科学界专门有一个学科叫“数值计算”，专门用来提升计算机进行各类计算时的准确性和效率问题。例如，著名的“梯度下降”以及“牛顿法”就是数值计算中的经典算法，也非常适合来处理求解函数极值的问题。梯度下降法是解决回归模型中最简单且有效的方法之一。

逻辑回归是一种与线性回归非常类似的算法，但是，从本质上讲，线性回归处理的问题类型与逻辑回归不一致。线性回归处理的是数值问题，也就是最后预测出的结果是数字，例如房价。而逻辑回归属于分类算法，也就是说，逻辑回归预测结果是离散的分类，例如判断这封邮件是否是垃圾邮件，以及用户是否会点击此广告链接等等。

在实现方面，逻辑回归只是对线性回归的计算结果加上了一个 Sigmoid 函数，将数值结果转化为了 0 到 1 之间的概率(Sigmoid 函数的图像一般来说并不直观，你只需要理解数值越大，函数越逼近 1；数值越小，函数越逼近 0)，接着我们根据这个概率可以做预测，例如概率大于 0.5，则这封邮件就是垃圾邮件，或者肿瘤是否是恶性的等等。从直观上来说，逻辑回归是画出了一条分类线，如图 5.18 所示。

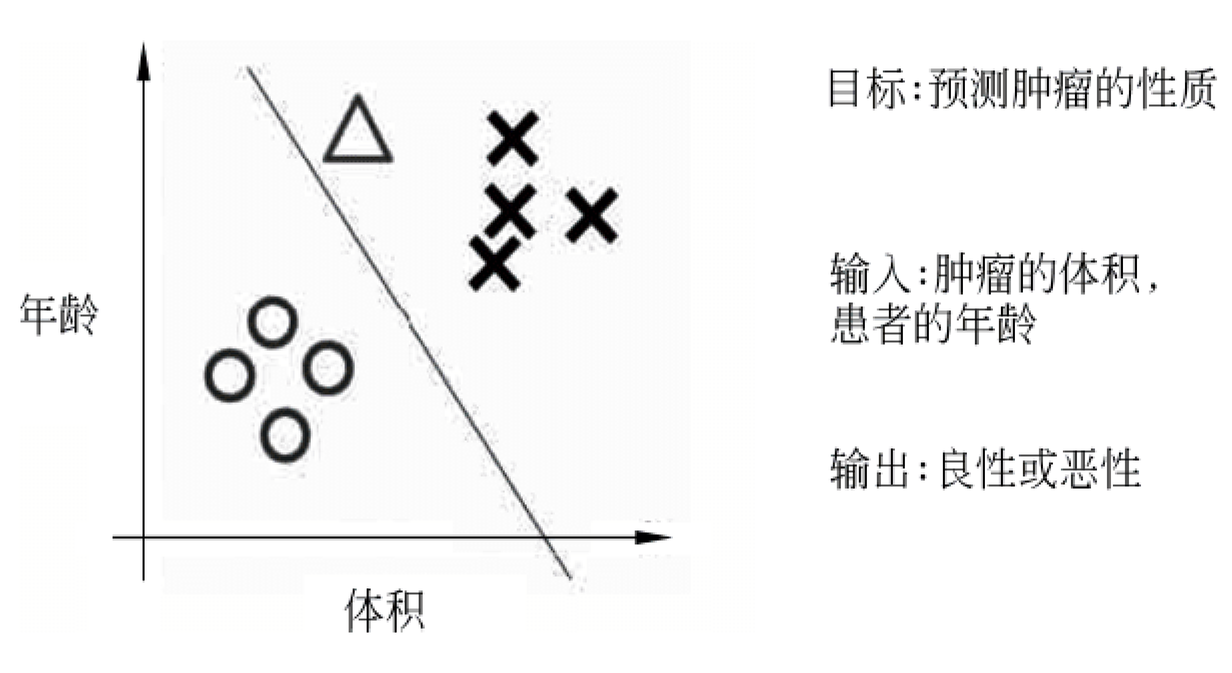


图 5.18 逻辑回归的直观解释

假设我们有一组肿瘤患者的数据，这些患者的肿瘤中有些是良性的(图中的○点)，有些是恶性的(图中的×点)。这里肿瘤的标志点(○点或×点)可以被称作数据的“标签”。同时每个数据包括两个“特征”：患者的年龄与肿瘤的大小。我们将这两个特征与标签映射到这个二维空间上，形成了图 5.18 中的数据。

当有一个绿色的点时，该判断这个肿瘤是恶性的还是良性的呢？根据标签点我们训练出了一个逻辑回归模型，也就是图中的分类线。这时，根据绿点出现在分类线的左侧，

因此我们判断它的标签应该是×，也就是说，属于恶性肿瘤。

逻辑回归算法划出的分类线基本都是线性的（也有划出非线性分类线的逻辑回归，不过那样的模型在处理数据量较大的时候效率会很低），这意味着当两类之间的界线不是线性时，逻辑回归的表达能力就不足。下面的两个算法是机器学习界最强大且重要的算法，都可以拟合出非线性的分类线。

2. 神经网络

神经网络（也称为人工神经网络，ANN）算法是 20 世纪 80 年代机器学习界非常流行的算法，不过在 20 世纪 90 年代中途衰落。现在，乘着“深度学习”之势，神经网络重装归来，重新成为最强大的机器学习算法之一。

神经网络的诞生起源于对大脑工作机理的研究。早期生物界学者们使用神经网络来模拟大脑。机器学习的学者们使用神经网络进行机器学习的实验，发现在视觉与语音的识别上效果都相当好。在 BP 算法（加速神经网络训练过程的数值算法）诞生以后，神经网络的发展形成了一股热潮。

具体说来，神经网络的学习机理是什么？简单来说，就是分解与整合。在著名的 Hubel-Wiesel 试验中，学者们研究猫的视觉分析机理就是这样的，如图 5.19 所示。

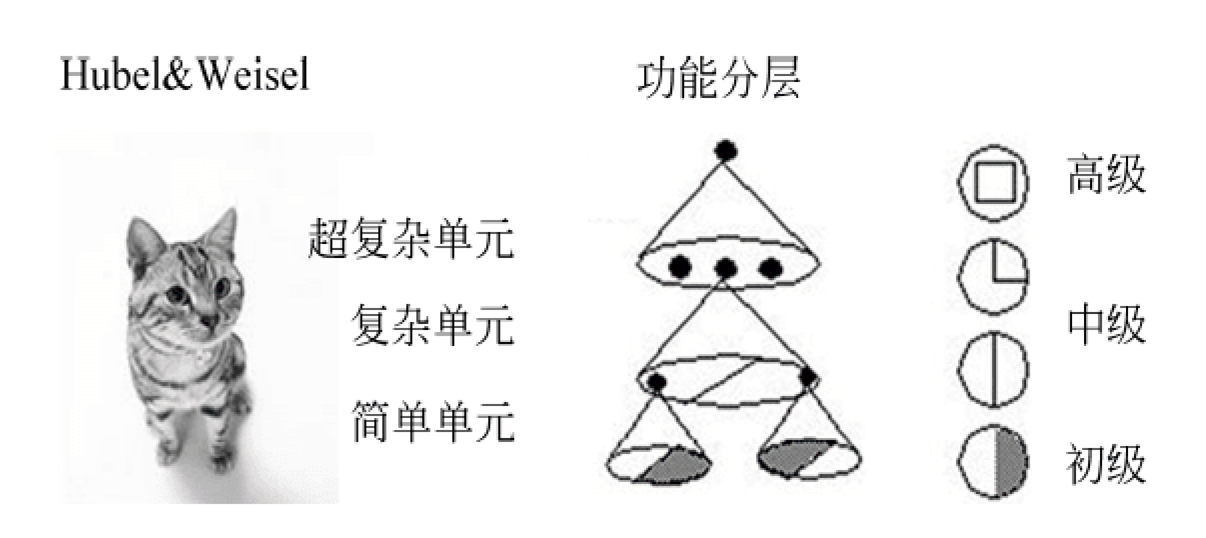


图 5.19　Hubel-Wiesel 试验与大脑视觉机理

比方说，一个正方形，分解为四个折线进入视觉处理的下一层中。四个神经元分别处理一个折线。每个折线再继续被分解为两条直线，每条直线再被分解为黑白两个面。于是，一个复杂的图像变成了大量的细节进入神经元，神经元处理以后再进行整合，最后得出了看到的是正方形的结论。这就是大脑视觉识别的机理，也是神经网络工作的机理。

让我们看一个简单的神经网络的逻辑架构。在这个网络中，分成输入层、隐藏层和输出层。输入层负责接收信号，隐藏层负责对数据的分解与处理，最后的结果被整合到输出层。每层中的一个圆代表一个处理单元，可以认为是模拟了一个神经元，若干个处理单元组成了一个层，若干个层再组成了一个网络，也就是“神经网络”。神经网络的逻辑架构如图 5.20 所示。

在神经网络中，每个处理单元事实上就是一个逻辑回归模型，逻辑回归模型接收上层的输入，把模型的预测结果作为输出传输到下一个层次。通过这样的过程，神经网络可以完成非常复杂的非线性分类。

图 5.21 演示了神经网络在图像识别领域的一个著名应用，这个程序叫做 LeNet，是

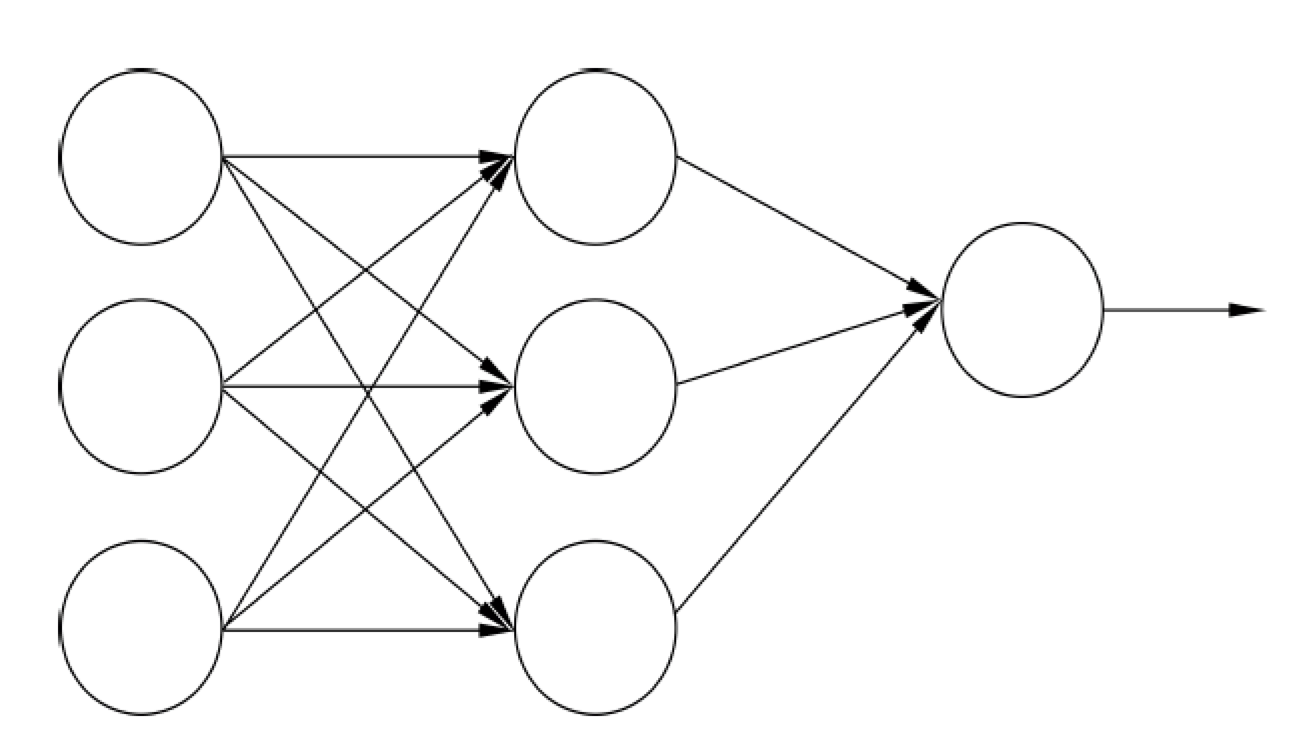


图 5.20　神经网络的逻辑架构

一个基于多个隐层构建的神经网络。通过 LeNet 可以识别多种手写数字,并且达到很高的识别精度与拥有较好的鲁棒性。

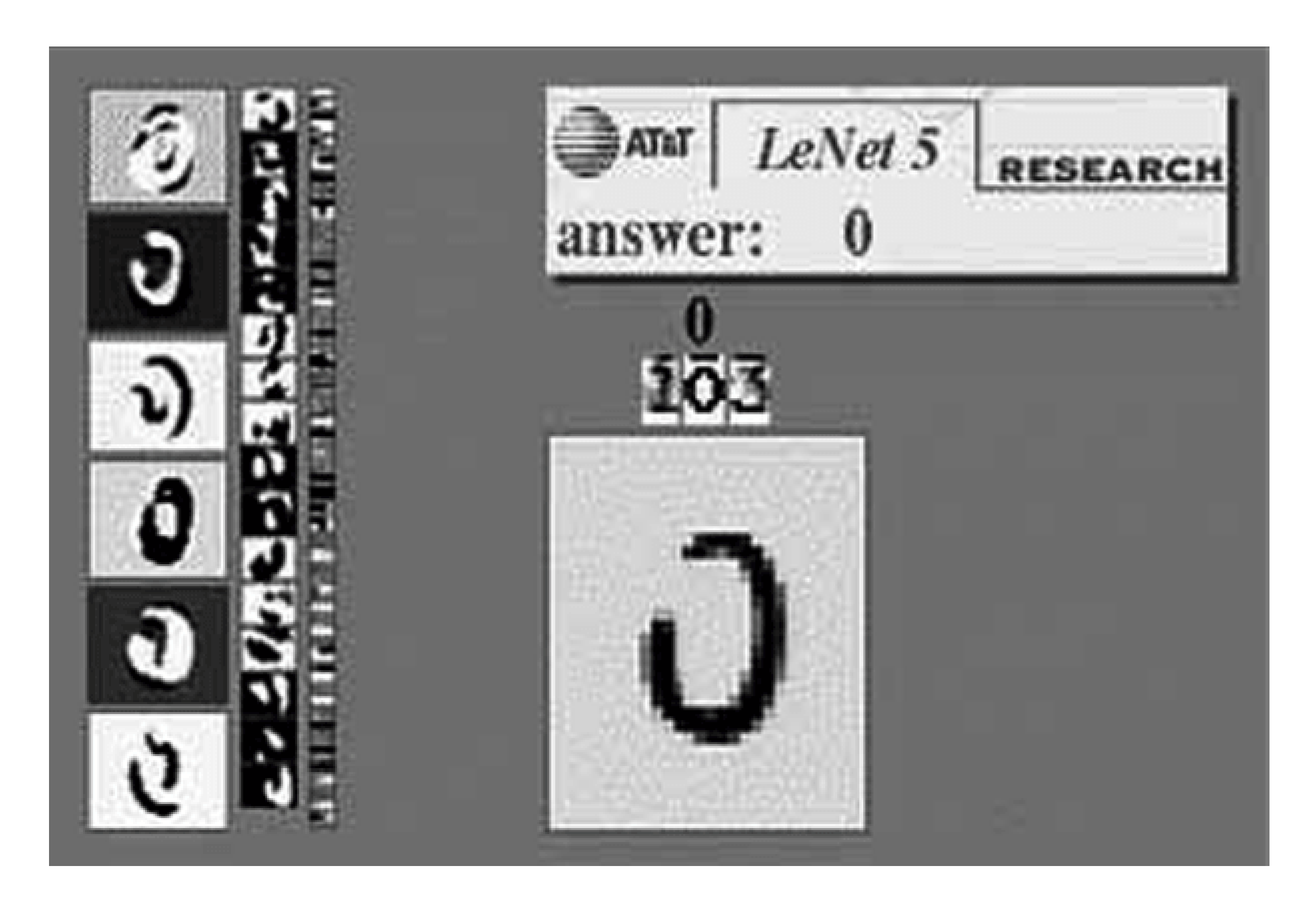


图 5.21　LeNet 的效果展示

图 5.21 右下方的方形中显示的是输入计算机的图像,方形上方的红色字样 answer 后面显示的是计算机的输出。左边的三条竖直的图像列显示的是神经网络中三个隐藏层的输出,可以看出:随着层次的不断深入,越深的层次处理的细节越低,例如层 3 基本处理的都已经是线的细节了。

进入 20 世纪 90 年代,神经网络的发展进入了一个瓶颈期。其主要原因是尽管有 BP 算法的加速,神经网络的训练过程仍然很困难。因此 20 世纪 90 年代后期支持向量机(SVM)算法取代了神经网络的地位。

3. SVM(支持向量机)

支持向量机算法是诞生于统计学习界,同时在机器学习界大放光彩的经典算法。

支持向量机算法从某种意义上来说是逻辑回归算法的强化:通过给予逻辑回归算法更严格的优化条件,支持向量机算法可以获得比逻辑回归更好的分类界线。但是如果没有某类函数技术,则支持向量机算法最多算是一种更好的线性分类技术。

但是,通过跟高斯"核"的结合,支持向量机可以表达出非常复杂的分类界线,从而取

得很好的分类效果。“核”事实上就是一种特殊的函数，最典型的特征就是可以将低维的空间映射到高维的空间，例如图5.22。

我们如何在二维平面划分出一个圆形的分类界线？在二维平面可能会很困难，但是通过“核”可以将二维空间映射到三维空间，然后使用一个线性平面就可以达成类似效果。也就是说，二维平面划分出的非线性分类界线可以等价于三维平面的线性分类界线。于是，我们可以通过在三维空间中进行简单的线性划分就可以取得在二维平面中的非线性划分效果。

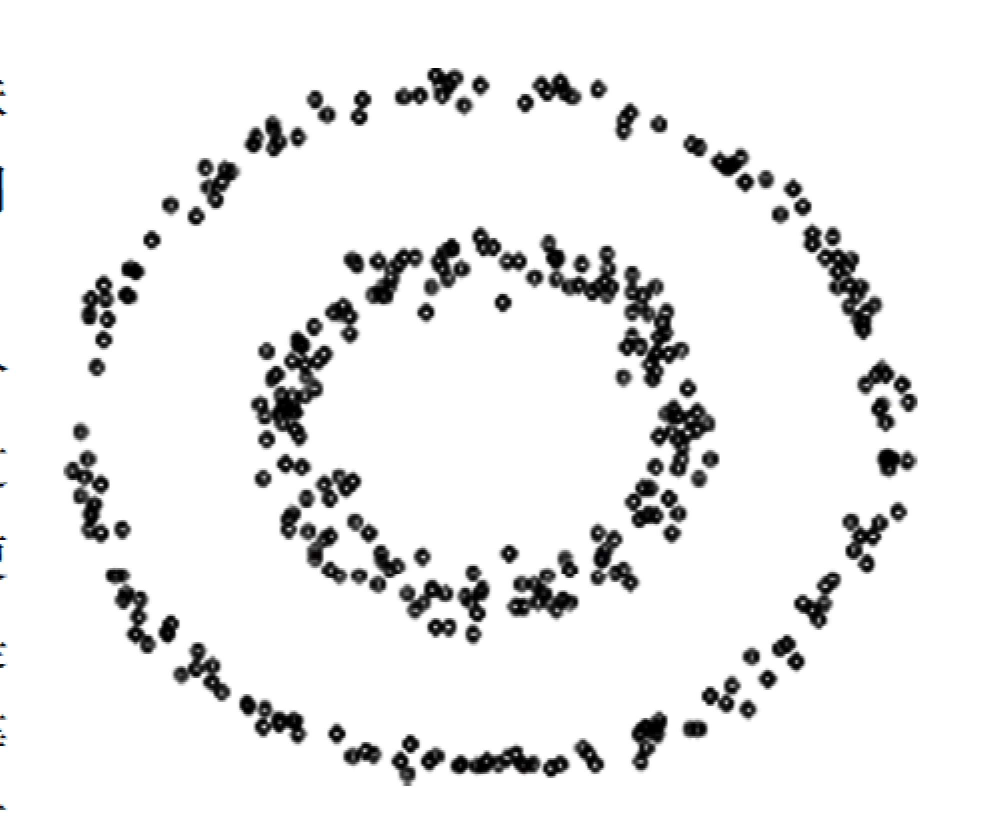


图5.22 支持向量机图例

支持向量机是一种偏数学的机器学习算法(相对的，神经网络则有生物科学成分)。在算法的核心步骤中，有一步证明，即将数据从低维映射到高维不会带来最后计算复杂性的提升。于是，通过支持向量机算法，既可以保持计算效率，又可以获得非常好的分类效果。因此支持向量机在20世纪90年代后期一直占据着机器学习中最核心的地位，基本取代了神经网络算法。直到现在神经网络借着深度学习重新兴起，两者之间才又发生了微妙的平衡转变。

4. 聚类算法

前面的算法中的一个显著特征就是训练数据中包含了标签，训练出的模型可以对其他未知数据预测标签。在下面的算法中，训练数据都是不含标签的，而算法的目的则是通过训练，推测出这些数据的标签。这类算法有一个统称，即无监督算法(前面有标签的数据的算法则是有监督算法)。无监督算法中最典型的代表就是聚类算法。

还是以一个二维的数据来说明，某一个数据包含两个特征。我希望通过聚类算法，给它们中不同的种类打上标签，我该怎么做呢？简单来说，聚类算法就是计算种群中的距离，根据距离的远近将数据划分为多个族群。

聚类算法中最典型的代表就是k-Means算法。

5. 降维算法

降维算法也是一种无监督学习算法，其主要特征是将数据从高维降低到低维层次。在这里，维度其实表示的是数据的特征量的大小，例如，房价包含房子的长、宽、面积与房间数量四个特征，也就是维度为四维的数据。可以看出，长与宽事实上与面积表示的信息重叠了，例如面积=长×宽。通过降维算法，就可以去除冗余信息，将特征减少为面积与房间数量两个特征，即从四维的数据压缩到二维。将数据从高维降低到低维，不仅利于表示，同时在计算上也能带来加速。

刚才说的降维过程中减少的维度属于肉眼可见的层次，同时压缩也不会带来信息的损失(因为信息冗余了)。如果肉眼不可见，或者没有冗余的特征，降维算法也能工作，不过这样会带来一些信息的损失。但是，降维算法可以从数学上证明，从高维压缩到的低维

中最大程度地保留了数据的信息。因此，使用降维算法仍然有很多的好处。

降维算法的主要作用是压缩数据与提升机器学习其他算法的效率。通过降维算法，可以将具有几千个特征的数据压缩至若干个特征。另外，降维算法的另一个好处是数据的可视化，例如将五维的数据压缩至二维，然后可以用二维平面来可视。降维算法的主要代表是 PCA 算法(即主成分分析算法)。

6. 推荐算法

推荐算法是目前业界非常流行的一种算法，在电商界，如亚马逊、天猫、京东等得到了广泛的运用。推荐算法的主要特征就是可以自动向用户推荐他们最感兴趣的东西，从而增加购买率，提升效益。推荐算法有两个主要的类别：

一类是基于物品内容的推荐，是将与用户购买的内容近似的物品推荐给用户，前提是每个物品都得有若干个标签，因此才可以找出与用户购买物品类似的物品，这样推荐的好处是关联程度较大，但是由于每个物品都需要贴标签，因此工作量较大。

另一类是基于用户相似度的推荐，则是将与目标用户兴趣相同的其他用户购买的东西推荐给目标用户，例如小 A 历史上买了物品 B 和 C，经过算法分析，发现另一个与小 A 近似的用户小 D 购买了物品 E，于是将物品 E 推荐给小 A。

两类推荐都有各自的优缺点，在电商应用中，一般是两类混合使用。推荐算法中最有名的算法就是协同过滤算法。

7. 其他

除了以上算法之外，机器学习界还有其他的如高斯判别、朴素贝叶斯、决策树等等算法。但是上面列的六个算法是使用最多、影响最广、种类最全的典型。机器学习界的一个特色就是算法众多，发展百花齐放。

下面做一个总结，按照训练的数据有无标签，可以将上面算法分为监督学习算法和无监督学习算法，但推荐算法较为特殊，既不属于监督学习，也不属于非监督学习，是单独的一类。

- 监督学习算法：线性回归、逻辑回归、神经网络、SVM。
- 无监督学习算法：聚类算法、降维算法。
- 特殊算法：推荐算法。

除了这些算法以外，有一些算法的名字在机器学习领域中也经常出现。但它们本身并不算是一个机器学习算法，而是为了解决某个子问题而诞生的。你可以将它理解为以上算法的子算法，用于大幅度提高训练过程。其中的代表有：梯度下降法，主要运用在线型回归、逻辑回归、神经网络、推荐算法中；牛顿法，主要运用在线性回归中；BP 算法，主要运用在神经网络中；SMO 算法，主要运用在 SVM 中。

5.4.4 机器学习的应用——大数据

说完机器学习的方法，下面谈一谈机器学习的应用。无疑，在 2010 年以前，机器学习的应用在某些特定领域发挥了巨大的作用，如车牌识别、网络攻击防范、手写字符识别等

等。但是,从 2010 年以后,随着大数据概念的兴起,机器学习大量的应用都与大数据高度耦合,几乎可以认为大数据是机器学习应用的最佳场景。

譬如,但凡你能找到的介绍大数据魔力的文章,都会说大数据如何准确预测到了某些事。例如经典的 Google 利用大数据预测了 H1N1 在美国某小镇的爆发,如图 5.23 所示。

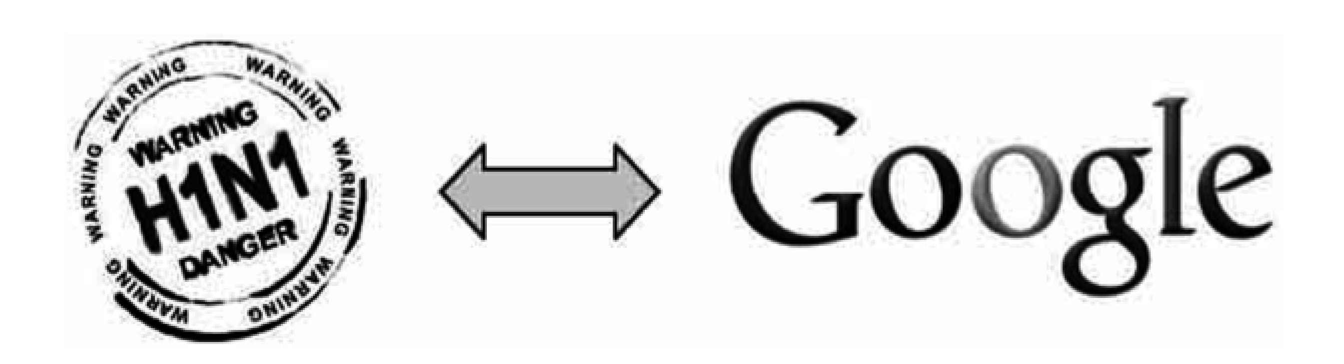


图 5.23　Google 成功预测 H1N1

百度预测 2014 年世界杯,从淘汰赛到决赛全部预测正确,如图 5.24 所示。



图 5.24　百度世界杯成功预测了所有比赛结果

这些实在太神奇了,那么究竟是什么原因导致大数据具有这些魔力的呢？简单来说,就是机器学习技术。正是基于机器学习技术的应用,数据才能发挥其魔力。

大数据的核心是利用数据的价值,机器学习是利用数据价值的关键技术,对于大数据而言,机器学习是不可或缺的。相反,对于机器学习而言,越多的数据越可能提升模型的精确性,同时,复杂的机器学习算法的计算时间也迫切需要分布式计算与内存计算这样的关键技术。因此,机器学习的兴盛也离不开大数据的帮助。大数据与机器学习两者是互相促进、相依相存的关系。

机器学习与大数据紧密联系。但是,必须清醒地认识到,大数据并不等同于机器学习,同理,机器学习也不等同于大数据。大数据中包含有分布式计算、内存数据库、多维分析等等多种技术。单从分析方法来看,大数据也包含以下四种分析方法：

(1) 大数据,小分析——即数据仓库领域的 OLAP 分析思路,也就是多维分析思想。

(2) 大数据,大分析——这个代表的就是数据挖掘与机器学习分析法。

(3) 流式分析——这个主要指的是事件驱动架构。

(4) 查询分析——经典代表是 NoSQL 数据库。

也就是说,机器学习仅仅是大数据分析中的一种而已。尽管机器学习的一些结果具有很大的魔力,在某种场合下是大数据价值最好的说明。但这并不代表机器学习是大数据下的唯一的分析方法。

机器学习与大数据的结合产生了巨大的价值。基于机器学习技术的发展,数据能够“预测”。对人类而言,积累的经验越丰富,阅历也广泛,对未来的判断越准确。例如常说的“经验丰富”的人比“初出茅庐”的人更有工作上的优势,就在于经验丰富的人获得的规律比他人更准确。而在机器学习领域,根据著名的一个实验,有效地证实了机器学习界一个理论:即机器学习模型的数据越多,机器学习的预测的效率就越好。

通过这张图可以看出,各种不同算法在输入的数据量达到一定级数后,都有相近的高准确度。于是诞生了机器学习界的名言:成功的机器学习应用不是拥有最好的算法,而是拥有最多的数据!

在大数据的时代,有好多优势促使机器学习能够应用更广泛。例如,随着物联网和移动设备的发展,我们拥有的数据越来越多,种类也包括图片、文本、视频等非结构化数据,这使得机器学习模型可以获得越来越多的数据。同时大数据技术中的分布式计算 MapReduce 使得机器学习的速度越来越快,可以更方便地使用。种种优势使得在大数据时代,机器学习的优势可以得到最佳的发挥。

5.4.5 机器学习的子类——深度学习

近来,机器学习的发展产生了一个新的方向,即“深度学习”。

深度学习的理念非常简单,就是传统的神经网络发展到了多隐藏层的情况。

20 世纪 90 年代以后,神经网络消寂了一段时间。但是 BP 算法的发明人 Geoffrey Hinton 一直没有放弃对神经网络的研究。由于神经网络在隐藏层扩大到两个以上,其训练速度就会非常慢,因此实用性一直低于支持向量机。2006 年,Geoffrey Hinton 在科学杂志 *Science* 上发表了一篇文章,论证了两个观点:

(1) 多隐层的神经网络具有优异的特征学习能力,学习得到的特征对数据有更本质的刻画,从而有利于可视化或分类;

(2) 深度神经网络在训练上的难度,可以通过“逐层初始化”来有效降低。

通过这样的发现,不仅解决了神经网络在计算上的难度,同时也说明了深层神经网络在学习上的优异性。从此,神经网络重新成为机器学习界中的主流强大学习技术。同时,具有多个隐藏层的神经网络被称为深度神经网络,基于深度神经网络的学习研究称为深度学习。

由于深度学习的重要性质,在各方面都取得了极大的关注,按照时间轴排序,有以下四个标志性事件值得一说:

2012年6月,《纽约时报》披露了Google Brain项目,这个项目是由Andrew Ng和MapReduce发明人Jeff Dean共同主导,用16 000个CPU Core的并行计算平台训练一种称为“深层神经网络”的机器学习模型,在语音识别和图像识别等领域获得了巨大的成功。Andrew Ng就是文章开始所介绍的机器学习的大牛(图5.25中右一立者)。





图5.25　Geoffrey Hinton与他的学生在Science上发表文章

2012年11月,微软在中国天津的一次活动上公开演示了一个全自动的同声传译系统,讲演者用英文演讲,后台的计算机一气呵成自动完成语音识别、英中机器翻译以及中文语音合成,效果非常流畅,其中支撑的关键技术是深度学习。

2013年1月,在百度的年会上,创始人兼CEO李彦宏高调宣布要成立百度研究院,其中第一个重点方向就是深度学习,并为此而成立深度学习实验室(IDL),如图5.26所示。



图5.26　深度学习的发展热潮

2013年4月,《麻省理工学院技术评论》杂志将深度学习列为2013年十大突破性技术(Breakthrough Technology)之首。

目前业界许多的图像识别技术与语音识别技术的进步都源于深度学习的发展,除了本文开头所提的Cortana等语音助手,还包括一些图像识别应用,其中典型的代表就是百度识图功能(见图5.27)。

深度学习属于机器学习的子类。基于深度学习的发展极大地促进了机器学习的地位提高,更进一步地,推动了业界对机器学习父类人工智能梦想的再次重视。

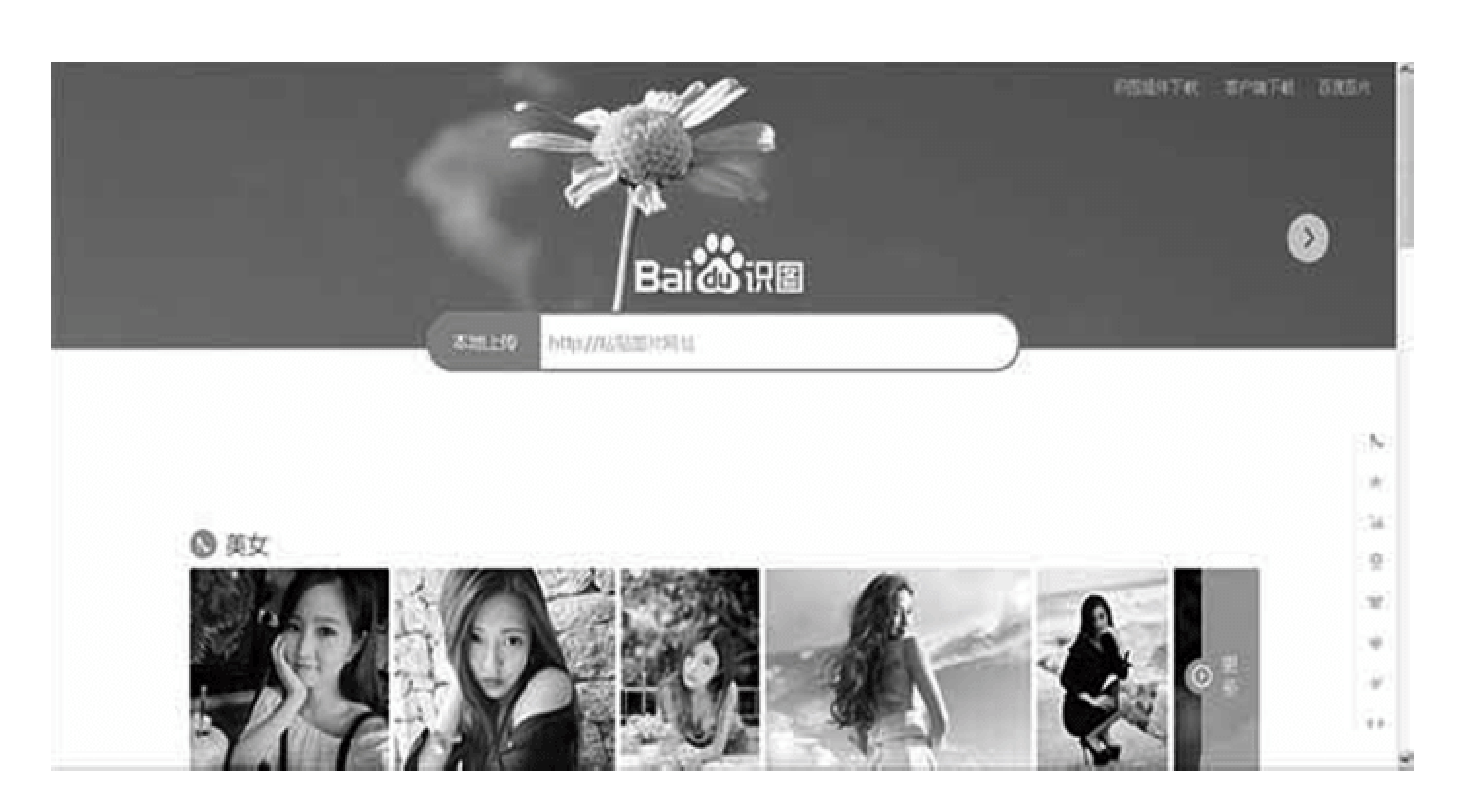


图 5.27 百度识图

5.4.6 机器学习的父类——人工智能

人工智能是机器学习的父类。深度学习则是机器学习的子类。三者的关系如图 5.28 所示。

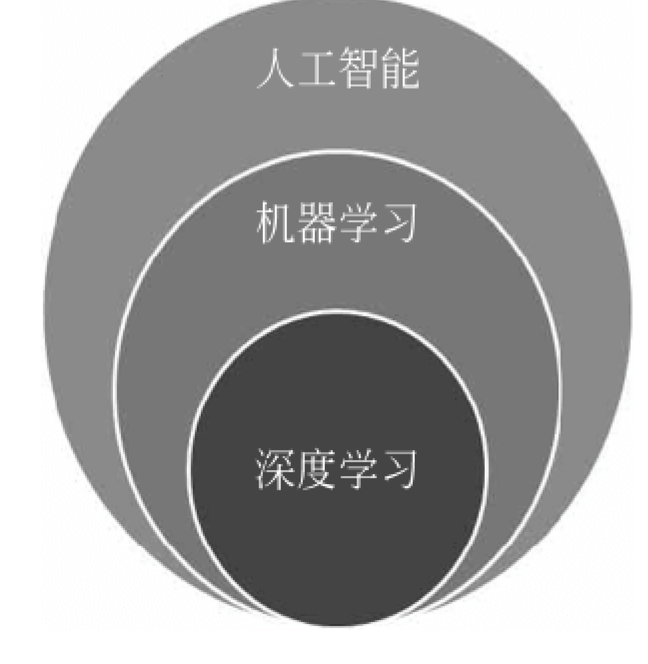


图 5.28 深度学习、机器学习、人工智能三者关系

毫无疑问，人工智能(AI)是人类所能想象的科技界最具突破性的发明了，某种意义上来说，人工智能就像游戏《最终幻想》的名字一样，是人类对于科技界的最终梦想。从 20 世纪 50 年代提出人工智能的理念以后，科技界，产业界不断在探索，研究。这段时间各种小说、电影都在以各种方式展现对于人工智能的想象。人类可以发明类似于人类的机器，这是多么伟大的一种理念！但事实上，自从 20 世纪 50 年代以后，人工智能的发展就不算顺利，未有见到足够震撼的科学技术的进步。

总结起来，人工智能的发展经历了如下若干阶段，从早期的逻辑推理，到中期的专家系统，这些科研进步确实使我们离机器的智能有点接近了，但还有一大段距离。直到机器学习诞生以后，人工智能界感觉终于找对了方向。基于机器学习的图像识别和语音识别在某些垂直领域达到了跟人相媲美的程度。机器学习使人类第一次如此接近人工智能的梦想。

事实上，如果我们把人工智能相关的技术以及其他业界的技术做一个类比，就可以发现机器学习在人工智能中的重要地位不是没有理由的。

人类区别于其他物体、植物、动物的最主要区别，作者认为是“智慧”。而智慧的最佳体现是什么？

是计算能力么，应该不是，心算速度快的人我们一般称之为天才。

是反应能力么，也不是，反应快的人我们称之为灵敏。

是记忆能力么，也不是，记忆好的人我们一般称之为过目不忘。

是推理能力么，这样的人我也许会称他智力很高，类似“福尔摩斯”，但不会称他拥有智慧。

是知识能力么，这样的人我们称之为博闻广，也不会称他拥有智慧。

想想看我们一般形容谁有大智慧？圣人，诸如庄子、老子等。智慧是对生活的感悟，是对人生的积淀与思考，这与我们机器学习的思想何其相似？通过经验获取规律，指导人生与未来。没有经验就没有智慧。

那么，从计算机来看，以上的种种能力都有种种技术去应对。

例如，计算能力我们有分布式计算，反应能力我们有事件驱动架构，检索能力我们有搜索引擎，知识存储能力我们有数据仓库，逻辑推理能力我们有专家系统，但是，唯有对应智慧中最显著特征的归纳与感悟能力，只有机器学习与之对应。这也是机器学习能力最能表征智慧的根本原因，如图 5.29 所示。

计算:云计算　　推理:专家系统

灵敏:事件驱动　　智慧:机器学习　　知识:数据仓库

检索:搜索引擎

图 5.29　机器学习与智慧

让我们再看一下机器人的制造，在具有了强大的计算、海量的存储、快速的检索、迅速的反应、优秀的逻辑推理后，如果再配合上一个强大的智慧大脑，一个真正意义上的人工智能也许就会诞生，这也是为什么说在机器学习快速发展的现在，人工智能可能不再是梦想的原因。

人工智能的发展可能不仅取决于机器学习，更取决于前面所介绍的深度学习，深度学习技术由于深度模拟了人类大脑的构成，在视觉识别与语音识别上显著性的突破了原有机器学习技术的界限，因此极有可能是真正实现人工智能梦想的关键技术。无论是 Google 大脑还是百度大脑，都是通过海量层次的深度学习网络所构成的。也许借助于深度学习技术，在不远的将来，一个具有人类智能的计算机真的有可能实现。

机器学习是目前业界最为 Amazing 与火热的一项技术，从网上的每一次淘宝的购买东西，到自动驾驶汽车技术，以及网络攻击抵御系统等等，都有机器学习的因子在其中，同时机器学习也是最有可能使人类完成 AI dream 的一项技术，各种人工智能目前的应用，如微软小冰聊天机器人，到计算机视觉技术的进步，都有机器学习努力的成分。作为一名当代计算机开发或管理人员，最好都应该了解一些机器学习的相关知识与概念，因为这可以帮你更好地理解为你带来莫大便利技术的背后原理，以及让你更好地理解当代科技的进程。

5.5　数据处理语言

5.5.1　数据分析语言 R

在 R 的官方教程里是这么给 R 下注解的：基 S 语言的一个数据分析和图形显示的

程序设计环境(A system for data analysis and visualization which is built based on S language)。

1. R的源起

原先AT&T贝尔实验室开发的一种用来进行数据探索、统计分析、作图的解释型语言——S语言,由John Chambers和同事开发,被用作一个统计分析平台。S是一种在编程环境操作的解释语言。S语法与C的语法很相似,但省去了困难的部分。S负责执行内存管理和变量声明,举例而言,这样用户就无须编写或调试这些方面了。更低的编程开销使得用户可以在同一个数据集上快速执行大量分析。

从一开始,S就考虑到了高级图形的创建,可向任何打开的图形窗口添加功能。可很容易地突出兴趣点,查询它们的值,使散点图变得更平滑,等等。

最初S语言的实现版本主要是S-PLUS。后来Auckland大学的Ross Ihaka和Robert Gentleman及其他志愿人员于1995年在S语言中创造了开源语言R,目的是专注于提供以更好和更人性化的方式做数据分析、统计和图形模型的语言。

开源语言R与S-PLUS有很多类似之处,两个软件有一定的兼容性。R是S的一种开源实现,是一种用于数据分析和图形的编程环境。

起初R主要是在学术和研究使用,但近来企业界发现R也很不错。这使得中的R成为企业中使用的全球发展最快的统计语言之一。

R的主要优势是它有一个庞大的社区,通过邮件列表、用户贡献的文档和一个非常活跃的堆栈溢出组提供支持。还有CRAN镜像,一个用户可以很简单地创造的一个包含R包的知识库。这些包有R里面的函数和数据,各地的镜像都是R网站的备份文件,完全一样,用户可以可以选择离你最近的镜像访问最新的技术和功能,而无须从头开发。

2. R是免费的

R是用于统计分析、绘图的语言和操作环境。R是一个自由、免费、源代码开放的软件,它是一个用于统计计算和统计制图的优秀工具。

R是一套完整的数据处理、计算和制图软件系统。其功能包括:数据存储和处理系统;数组运算工具(其向量、矩阵运算方面功能尤其强大);完整连贯的统计分析工具;优秀的统计制图功能;简便而强大的编程语言:可操纵数据的输入和输入,可实现分支、循环,用户可自定义功能。

R是一个免费的自由软件,它有UNIX、Linux、Mac OS和Windows版本,都是可以免费下载和使用的,在那儿可以下载到R的安装程序、各种外挂程序和文档。在R的安装程序中只包含了8个基础模块,其他外在模块可以通过CRAN获得。

3. R的特点

(1) 有效的数据处理和保存机制。

(2) 拥有一整套数组和矩阵的操作运算符。

(3) 一系列连贯而又完整的数据分析中间工具。

(4) 图形统计可以对数据直接进行分析和显示,可用于多种图形设备。

（5）一种相当完善、简洁和高效的程序设计语言。它包括条件语句、循环语句、用户自定义的递归函数以及输入输出接口。

（6）R 语言是彻底面向对象的统计编程语言。

（7）R 语言和其他编程语言、数据库之间有很好的接口。

（8）R 语言是自由软件，可以放心大胆地使用，但其功能却不比任何其他同类软件差。

（9）R 语言具有丰富的网上资源。

4. 做数据分析必须学 R 的理由

R 是一种灵活的编程语言，专为促进探索性数据分析、经典统计学测试和高级图形学而设计。R 拥有丰富的、仍在不断扩大的数据包库，处于统计学、数据分析和数据挖掘发展的前沿。R 已证明自己是不断成长的大数据领域的一个有用工具，并且已集成到多个商用包中，比如 IBM SPSS®、InfoSphere® 以及 Mathematica。

5.5.2　大数据开发语言 Python

Python 是一种面向对象、直译式的计算机程序语言，具有近二十年的发展历史。它包含了一组功能完备的标准库，能够轻松完成很多常见的任务。它的语法简单，与其他大多数程序设计语言使用大括号不一样，它使用缩进来定义语句块。

Python 具备垃圾回收功能，能够自动管理内存使用。它经常被当作脚本语言用于处理系统管理任务和网络程序编写，然而它也非常适合完成各种高级任务。Python 虚拟机本身几乎可以在所有的作业系统中运行。使用一些诸如 py2exe、PyPy、PyInstaller 之类的工具可以将 Python 源代码转换成可以脱离 Python 解释器运行的程序。

Python 的官方解释器是 CPython，该解释器用 C 语言编写，是一个由社区驱动的自由软件，目前由 Python 软件基金会管理。

Python 支持命令式程序设计、面向对象程序设计、函数式编程、面向侧面的程序设计、泛型编程多种编程范式。

1. 大数据全栈式开发语言——Python

只要会 JavaScript 就可以写出完整的 Web 应用，只要会 Python，就可以实现一个完整的大数据处理平台。表 5.1 给出了 Python 的应用领域。

表 5.1　Python 应用领域

领　　域	流行语言	领　　域	流行语言
云基础设施	Python，Java，Go	网络爬虫	Python，PHP，C++
DevOps	Python，Shell，Ruby，Go	数据处理	Python，R，Scala

在理论研究领域，R 语言也许是最受数据科学家欢迎的，但是 R 语言的问题也很明显，因为是统计学家们创建了 R 语言，所以其语法略显怪异。而且 R 语言要想实现大规模分布式系统，还有很长一段时间的工程之路要走。所以很多公司使用 R 语言做原型试

验，算法确定之后，再翻译成工程语言。

Python 也是数据科学家最喜欢的语言之一。和 R 语言不同，Python 本身就是一门工程性语言，数据科学家用 Python 实现的算法，可以直接用在产品中，这对于大数据初创公司节省成本是非常有帮助的。正是因为数据科学家对 Python 和 R 的热爱，Spark 为了讨好数据科学家，对这两种语言提供了非常好的支持。

Python 的数据处理相关类库非常多。高性能的科学计算类库 NumPy 和 SciPy，给其他高级算法奠定了非常好的基础，matploglib 让 Python 画图变得像 Matlab 一样简单。Scikit-learn 和 Milk 实现了很多机器学习算法，基于这两个库实现的 Pylearn2，是深度学习领域的重要成员。Theano 利用 GPU 加速，实现了高性能数学符号计算和多维矩阵计算。当然，还有 Pandas，一个在工程领域已经广泛使用的大数据处理类库，其 DataFrame 的设计借鉴自 R 语言，后来又启发 Spark 项目实现类似机制。

2. 为什么是 Python

正是因为应用开发工程师、运维工程师、数据科学家都喜欢 Python，才使得 Python 成为大数据系统的全栈式开发语言。

(1) 对于开发工程师而言，Python 的优雅和简洁无疑是最大的吸引力，在 Python 交互式环境中，执行 import this，读一读 Python 之禅，你就明白 Python 为什么如此吸引人。Python 社区一直非常有活力，和 NodeJS 社区软件包爆炸式增长不同，Python 的软件包增长速度一直比较稳定，同时软件包的质量也相对较高。有很多人诟病 Python 对于空格的要求过于苛刻，但正是因为这个要求，才使得 Python 在做大型项目时比其他语言有优势。OpenStack 项目总共超过 200 万行代码，证明了这一点。

(2) 对于运维工程师而言，Python 的最大优势在于，几乎所有 Linux 发行版都内置了 Python 解释器。Shell 虽然功能强大，但毕竟语法不够优雅，写比较复杂的任务会很痛苦。用 Python 替代 Shell，做一些复杂的任务，对运维人员来说，是一次解放。

(3) 对于数据科学家而言，Python 简单又不失强大。和 C/C++ 相比，不用做很多底层工作，可以快速进行模型验证；和 Java 相比，Python 语法简洁，表达能力强，同样的工作只需要 1/3 代码；和 Matlab、Octave 相比，Python 的工程成熟度更高。不止一个编程大牛表达过，Python 是最适合作为大学计算机科学编程课程使用的语言——MIT 的计算机入门课程就是使用的 Python——因为 Python 能够让人学到编程最重要的东西——如何解决问题。

顺便提一句，微软参加 2015 年 PyCon，高调宣布提高 Python 在 Windows 上的编程体验，包括 Visual Studio 支持 Python，优化 Python 的 C 扩展在 Windows 上的编译等等。

3. R 和 Python 的区别

1) R 和 Python：数字的比较

在网上可以经常看到比较 R 和 Python 人气的数字，虽然这些数字往往就这两种语言是如何在计算机科学的整体生态系统不断发展，但是很难并列进行比较。主要的原因是，R 仅在数据科学的环境中使用，而 Python 作为一种通用语言，被广泛应用于许多领域，如网络的发展。这往往导致排名结果偏向于 Python，而且从业者工资会较低。

2）如何使用 R

R 主要用于当数据分析任务需要独立的计算或分析单个服务器。这是探索性的工作,因为 R 有很多包和随时可用的测试,可以提供必要的工具,快速启动和运行的数量庞大几乎任何类型的数据分析。R 甚至可以是一个大数据解决方案的一部分。

3）如何使用 Python

如果你的数据分析任务需要使用 Web 应用程序,或代码的统计数据需要被纳入生产数据库进行集成时可以使用 Python,作为一个完全成熟的编程语言,它是实现算法一个伟大的工具。

4）R 和 Python：数据科学行业的表现

如果你看一下最近的民意调查,在数据分析的编程语言方面,R 是明显的赢家。

有越来越多的人从研发转向 Python。此外,有越来越多的公司使用这两种语言来进行组合。

如果你打算从事数据行业,你用好学会这两种语言。招聘趋势显示这两个技能的需求日益增加,而工资远高于平均水平。

最终你该学习什么呢：

由你决定！作为一个数据工作者,你需要在工作中选择最适合需要的语言。在学习之前问清楚这些问题可以帮助你：

你想解决什么问题？

什么是学习语言的净成本？

是什么在你的领域中常用的工具？

什么是其他可用工具以及如何做这些涉及的常用工具？

5.6　大数据应用案例之：北京的人流在哪儿？用大数据看城市

如何读懂一座城市？人们把生活构建在大大小小的城市中,城市不仅为人们提供工作机会,更寄托着休闲、娱乐、教育等诸多期待。在这个复杂的网络、动态的系统之中,每个人只能看到自己周围的生活,而几乎无法了解整个城市的场景。尤其是,如果你生活在一个特大城市,比如常住人口超过 2300 万的北京,可能穷尽一生都无法彻底读懂这座被尊称为帝都的城市。

如今,我们有了“大数据”这样的信息时代新利器,每日都能直观俯视城市日新月异的变化,不必只从平面地图和县志中来间接理解城市。

毕竟,房子和土地只是表象,人的聚集才是城市的本质。就像使用卫星地图监控城市的土地开发那样,我们现在利用大数据,在不同层次监测人口聚集,更好地回答“人在哪儿”的基本问题。

1. 传统的宏观统计

以前我们只能看到宏观统计,例如采用县级统计年鉴数据库分析全国尺度的区县域人口密度(2012 年),宏观表现全国人口分布的京津冀、珠三角、长三角和成渝经济圈四极

大结构。

如果把尺度放得更小一些，我们又能看到什么？我们采用街道尺度的第六次人口普查数据，分析了北京市域街道层面的人口总量和人口密度分布（乡镇街道立体图中，高度和颜色深浅度分别表示人口的数量和密度），如图 5.30 所示。

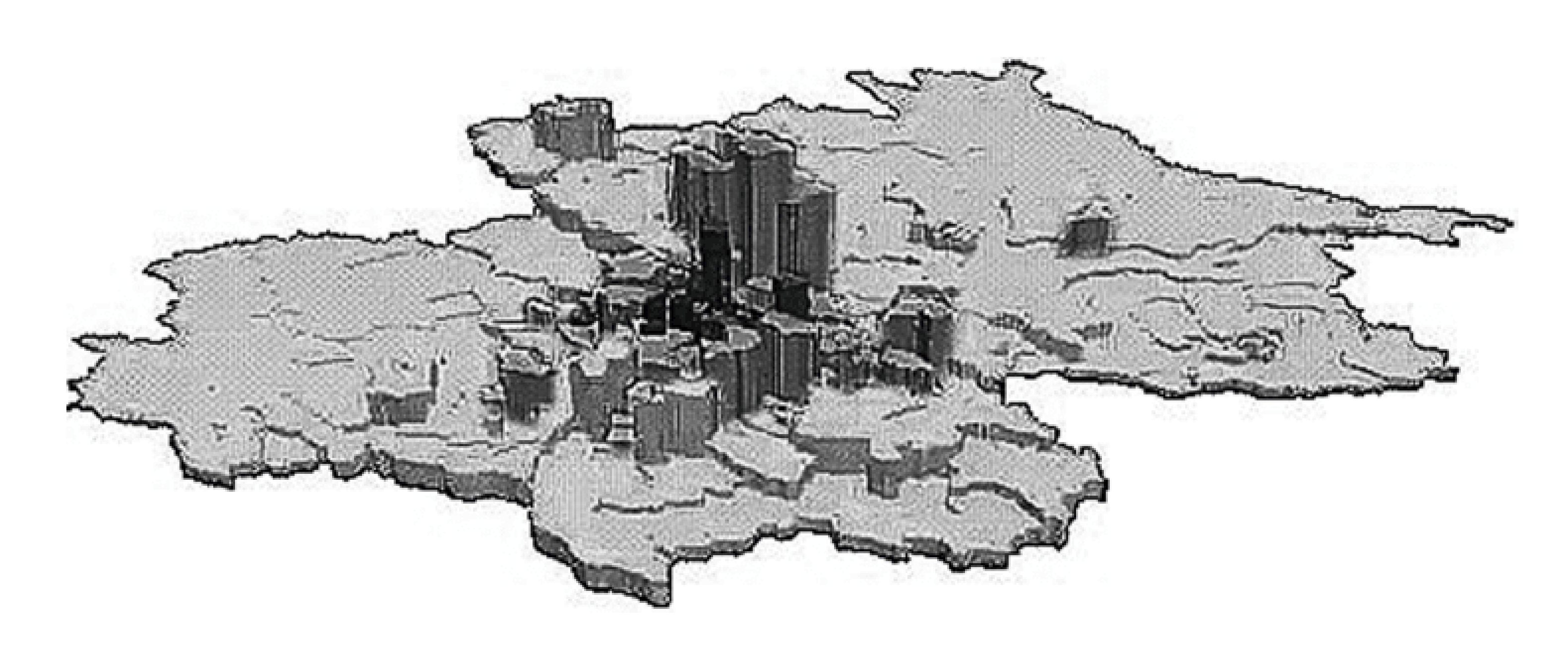


图 5.30 北京市域人口总量和人口密度分布

从人口总量看，昌平区的回龙观、东小口镇（天通苑）、北七家镇（天通苑以北），海淀区的学院路、北太平庄街道，以及大兴区的黄村地区，都聚集了大量人口；而从人口密度看，高密度区主要集中在海淀区和西城区。因聚集了大量的优质教育资源，海淀区在总量和密度上均呈现较高的值，所谓“宇宙中心”，果然不虚。

用大数据回答“人在哪儿”的问题。

上述数据可以让我们了解城市的脉络，但从中终究无法看到时间如何在城市中流逝、人们在城市中如何运动。由此，我们在这里尝试用大数据去回答城市中“人在哪儿”，把时间维度放进城市空间分析，重新理解城市中人的活动。

2. 北京：在哪儿上班，在哪儿睡觉

我们采用百度（百度热力图）和腾讯（宜出行平台）实时网格人口数据，选择工作日上午 10 点和夜间 23 点，分别代表上班工作和下班居家的活动状态，由此得出城市的职住中心。

就业中心主要集中在中关村、知春路、朝阳门-建国门-国贸一带、王府井-东单、金融街、西单、西直门、上地、望京、东直门、亮马桥、朝阳路十里堡段、惠新西街南北口、五道口、六道口等（北京南站因处于交通枢纽而聚集较多人群）。

按照夜间 23 点的人口分布（即居住分布）情况，可以发现，居住中心主要集中在中关村、回龙观、西小口、六道口、五道口、牡丹园、清河、知春路、大钟寺、学院南路、劲松-潘家园、宋家庄-石榴庄、京沪高速与南六环相交处、十里堡、望京、北苑、立水桥、天通苑、芍药居、小营等地。

通过对比，可以发现城市白天和黑夜的不同形态。第一种空间，白天熙熙攘攘的金融街、国贸、西单、王府井等商业就业中心，到了晚上一片寂静；第二种空间，集商业、就业、居住于一体的中关村、五道口、六道口、知春路等地，无论白天黑夜均集聚大量人气；第三种空间，回龙观、天通苑、北苑、宋家庄等主要以居住为主的地区，体现了睡城的基本特征。由此，大数据可以帮助我们了解城市居民如何使用城市空间，进行实时动态监测。

奥林匹克森林公园南园：哪里人多？哪里人少？

大数据不光能识别宏观的职住分布，还可以分析微观的公共空间，如小区公园、购物商场的使用情况。

同样，我们采用百度景区热图数据，配以实时人流动画作为表现形式，便可得出人们对微观空间的使用情况。例如，由清华同衡规划设计研究院主持规划设计的奥林匹克森林公园南园，除了地铁站森林公园南门以外，人流主要沿5km的规划环道分布，到了晚上表现尤为明显，而3km的规划环道上并未形成明显的人流集聚。同时，在以南门、西门和东门为核心的周边区域，有部分人流集聚，仰山所在中心区域则明显十分稀疏。

如图5.31所示，对比工作日周五和周末周六的人流量，可以发现，周五的人流量主要集中在外围5km的环道，周六的人群分布则更为广泛，更加深入到奥森公园内部的各处景点。同时对比早间和晚间的人流量，周五早上8:30的人流量要明显高于夜晚20:30，而在周六早上和夜晚的人流量差异较小。

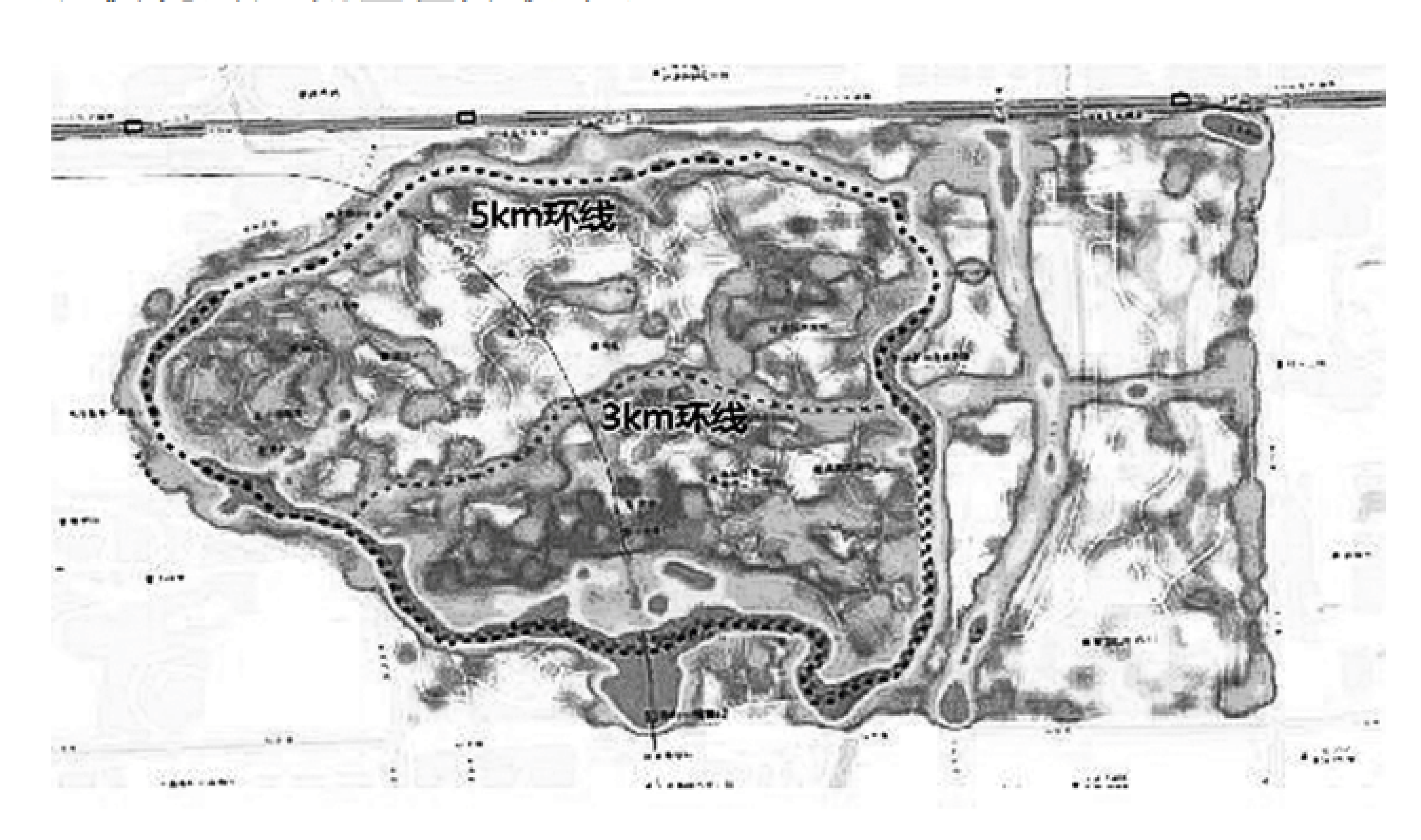


图5.31　百度景区热图数据

其实设计者也无法准确预测到这些现实使用情况，这不禁让人思考，是否3km的规划环道以及仰山所在中心区域的设施配套不足，导致使用率低？5km环道，是否因塑胶跑道而吸引了大部分人群；而仰山所在中心区域因灯光昏暗，且较少道路连通园门口，所以人群较少？这可能是设计师的精心安排，但有些可能是疏忽。在规划实施评估和未来的规划改进中，可以有针对性地进行优化。这是利用大数据发掘微观尺度空间使用模式的例子。

三里屯太古里和SOHO，各自的商业特征如何？

三里屯太古里与三里屯SOHO均处三里屯核心地段，在地理位置上几无优劣之分，仅隔一条工体北路。但人们可能会有一种体会：SOHO门可罗雀，太古里时尚繁华。

但实情和观感一致吗？暂把7月14日优衣库事件对人流的影响放在一边，通过图表(数据来自腾讯宜出行平台和百度景区热力图，人口数值经技术处理，不完全代表人数)，可以发现，无论工作日还是周末，三里屯SOHO人流量均高于太古里人流量。当然，可以在图里清楚读到太古里在事件后收获的人流增量，如图5.32所示。

我们只感受到太古里川流不息的观光购物人潮，却没有看到SOHO的高楼里“藏匿”的上班族和住客。

从用地性质看，太古里和SOHO均属商业用地，规划图纸上标注的是同一种颜色，但它们真的一样吗？从具体使用功能上，我们发现，太古里和SOHO运营的其实是不同类

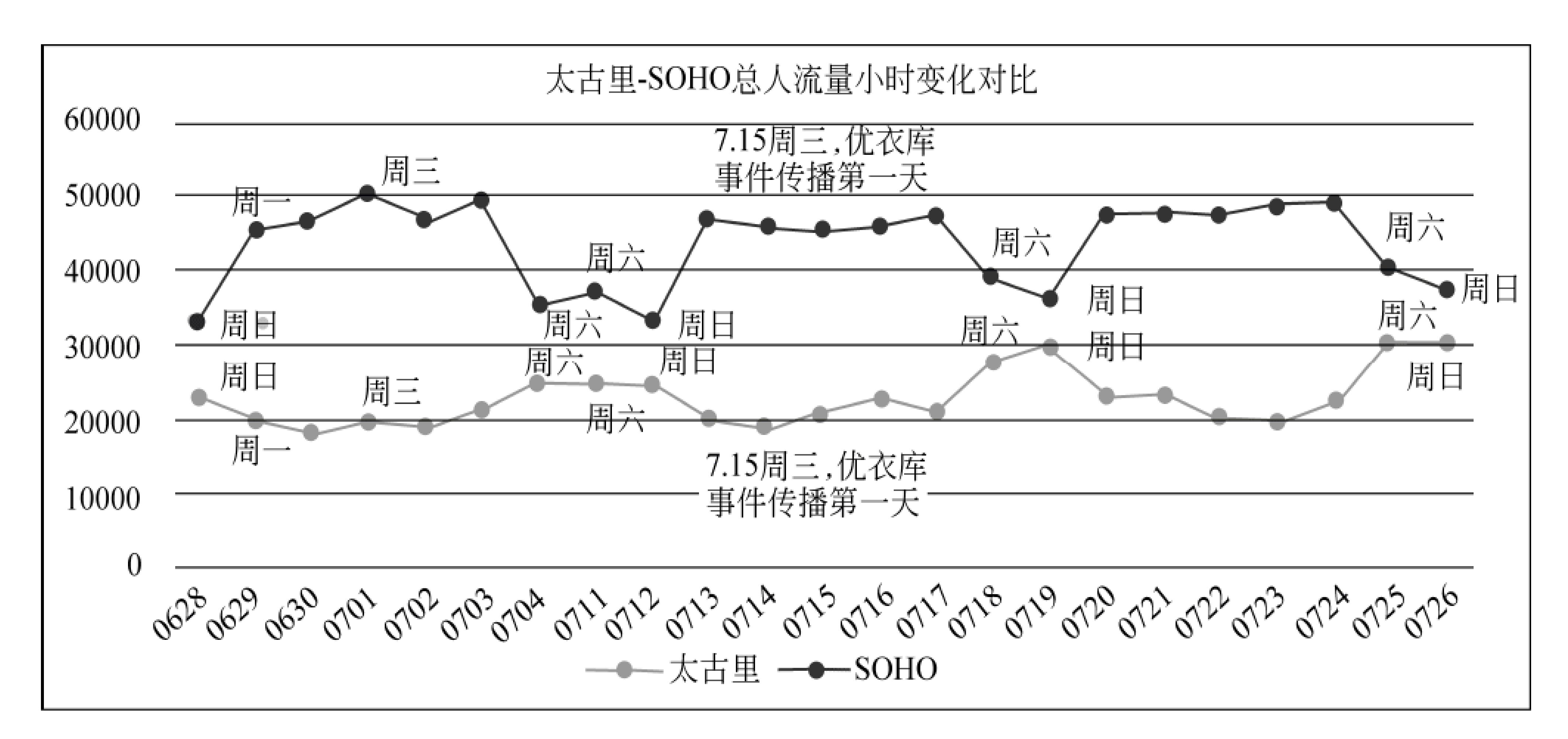


图 5.32 太古里和 SOHO 总人流量小时变化对比

型的商业项目。太古里定位是综合休闲娱乐区,是以开放式购物区为主的商业综合体;而SOHO 则是集商业、办公、居住为一体的综合社区。这两个地块随时间变化的人流量曲线,体现了它们承担功能的差异。

在太古里,周末人流量大于工作日;而日间人流量随时间推移缓慢增多,午间 13:00 左右增至最高峰,晚间 22:00 点之后,人群逐渐散尽。所以,太古里是人们休闲娱乐的去处,夜幕降临后大家各回各家。而在 SOHO,与太古里相反,工作日人流量大于周末;SOHO 达到人流量高峰的时间段,也比太古里提前,在早间 10:00 左右攀至最高峰,这正是上班族们陆续到达单位打卡开始一天工作的时间段。

可见,精细化的实时网络数据能精确刻画不同使用模式地块的人口时空特征,是我们厘清复杂城市系统线团的一根解锁线头。当然,以上分析只是简单示意,我们还将使用机器学习等技术对其进行更深入的分析和建模,以及实践应用。

3. 总结

基于上述三个用大数据进行的“人在哪儿”(全北京职住分布、奥林匹克森林公园南园人流实时变化动图、三里屯太古里和 SOHO 人口曲线特征刻画)的分析,我们的城市从二维的地图和文字中“活”了起来。

我们可以观察到城市在全天 24 小时的不同面貌,人流在公园等公共空间如何聚集,纯商业项目和综合社区在不同时间段以及时间点人流量的差异。

如果说,传统统计数据特征是平面、静态和粗放的,那么大数据则让城市的数据维度走向立体、动态和精确。如果说传统的统计数据主要服务于执政者从上至下的行政管理,那么大数据则服务于自下而上的问题解决。

大城市人地矛盾的确已十分突出,政府政策制定时往往首先想到疏解人口。但事实上,依靠数据提升精细化的规划和管理水平后,我们的城市也可以和东京等城市一样,更好地满足不同人群的基础设施和公共服务需求,最大化发挥有限设施的服务水平,提高其使用效率。可以说,大数据让城市和生活更加融合,让空间和市民更加贴近,最终能让我们的城市生活更加美好。

习题与思考题

一、选择题

1. 大数据与三个重大的思维转变有关，这三个转变是什么？（　　）（多选题）

A. 要分析与某事物相关的所有数据，而不是依靠分析少量的数据样本

B. 我们乐于接受数据的纷繁复杂，而不再追求精确性

C. 在数字化时代，数据处理变得更加容易、更加快速，人们能够在瞬间处理成千上万的数据

D. 我们的思想发生了转变，不再探求难以捉摸的因果关系，转而关注事物的相关关系

2. 下面关于大数据的解说正确的是（　　）。（多选题）

A. 大数据是人们在大规模数据的基础上可以做到的事情，而这些事情在小规模数据的基础上是无法完成的

B. 大数据是人们获得新的认知、创造新的价值的源泉

C. 大数据还是改变市场、组织机构以及政府与公民关系的方法

D. 无效的数据越来越多

3. 大数据的科学价值和社会价值正是体现在（　　）。（多选题）

A. 一方面，对大数据的掌握程度可以转化为经济价值的来源

B. 另一方面，大数据已经撼动了世界的方方面面，从商业科技到医疗、政府、教育、经济、人文以及社会的其他各个领域

C. 大数据的价值不再单纯来源于它的基本用途，而更多源于它的二次利用

D. 大数据时代，很多数据在收集的时候并无意用作其他用途，而最终却产生了很多创新性的用途

4. 关于大数据的概念正确的有（　　）。（多选题）

A. 大数据时代要求我们重新审视精确性的优劣

B. 大数据不仅让我们不再期待精确性，也让我们无法实现精确性

C. 错误并不是大数据固有的特性，而是一个亟须我们去处理的现实问题，并且有可能长期存在

D. 错误性是大数据本身固有的

5. 社会将两个折中的想法不知不觉地渗入了我们的处事方法中，我们甚至不再把这当成一种折中，而是把它当成了事物的自然状态。这两个折中的方法是什么？（　　）（多选题）

A. 第一个折中是我们默认自己不能使用更多的数据，所以我们就不会去使用更多的数据

B. 第二个折中出现在数据的质量上

C. 第一个折中是我们能够容忍模糊和不确定出现在一些过去依赖于清晰和精确的领域

D. 第二个折中是能够得到一个事物更完整的概念,我们就能接受模糊和不确定的存在

6. 数据化最早的根基是什么?(　　)。(多选题)

A. 计量　　B. 数字化　　C. 记录　　D. 阿拉伯数字

7. 关于数据的潜在价值,说法正确的是(　　)。(多选题)

A. 数据的真实价值就像漂浮在海洋中的冰山,第一眼只能看到冰山一角,而绝大部分则隐藏在表面之下

B. 判断数据的价值需要考虑到未来它可能被使用的各种方式,而非仅仅考虑其目前的用途

C. 在基本用途完成后,数据的价值仍然存在,只是处于休眠状态

D. 数据的价值是其所有可能用途的总和

8. MapReduce 的 Map 函数产生很多的(　　)。

A. key　　B. value

C. <key,value>　　D. Hash

9. Page Rank 是一个函数,它对 Web 中的每个网页赋予一个实数值。它的意图在于网页的 Page Rank 越高,那么它就(　　)。

A. 相关性越高　　B. 越不重要　　C. 相关性越低　　D. 越重要

10. 大数据的简单算法与小数据的复杂算法相比(　　)。

A. 更有效　　B. 相当　　C. 不具备可比性　　D. 无效

二、问答题

1. 什么是实时交互计算?什么是流计算?
2. 请解释数据分类与聚类的概念。
3. 什么是数据集成?
4. 请详述机器学习的定义和例子。
5. 请概述机器学习在大数据方面的应用。
6. 解释数据分析语言 R 和大数据开发语言 Python 的区别。

第 6 章　大数据查询、显现与交互

6.1　数据的查询

6.1.1　常规数据库查询结构化数据

数据库是为便于有效地管理信息而创建的，人们希望数据库可以随时提供所需要的数据信息。因此，对用户来说，数据查询是数据库最重要的功能。在数据库中创建了对象并且在基表中添加了数据后，用户便可以从数据库中检索特定信息。

结构化查询语言(Structured Query Language)是一种特殊目的的编程语言，是一种数据库查询和程序设计语言，用于存取数据以及查询、更新和管理关系数据库系统；同时也是数据库脚本文件的扩展名。

结构化查询语言是高级的非过程化编程语言，允许用户在高层数据结构上工作。它不要求用户指定对数据的存放方法，也不需要用户了解具体的数据存放方式，所以具有完全不同底层结构的不同数据库系统，可以使用相同的结构化查询语言作为数据输入与管理的接口。结构化查询语言语句可以嵌套，这使它具有极大的灵活性和强大的功能。

1986 年 10 月，美国国家标准协会对 SQL 进行规范后，以此作为关系式数据库管理系统的标准语言(ANSI X3. 135-1986)，1987 年得到国际标准组织的支持下成为国际标准。不过各种通行的数据库系统在其实践过程中都对 SQL 规范做了某些编改和扩充。所以，实际上不同数据库系统之间的 SQL 不能完全相互通用。

结构化查询语言包含 6 个部分。

1. 数据查询语言(Data Query Language，DQL)

其语句也称为“数据检索语句”，用于从表中获得数据，确定数据怎样在应用程序给出。保留字 SELECT 是 DQL(也是所有 SQL)用得最多的动词，其他 DQL 常用的保留字有 WHERE、ORDER BY、GROUP BY 和 HAVING。这些 DQL 保留字常与其他类型的 SQL 语句一起使用。

2. 数据操作语言(Data Manipulation Language，DML)

其语句包括动词 INSERT、UPDATE 和 DELETE，它们分别用于添加、修改和删除表中的行，也称为动作查询语言。

3. 事务处理语言(TPL)

它的语句能确保被 DML 语句影响的表的所有行及时得以更新。TPL 语句包括 BEGIN TRANSACTION、COMMIT 和 ROLLBACK。

4. 数据控制语言(DCL)

它的语句通过 GRANT 或 REVOKE 获得许可，确定单个用户和用户组对数据库对象的访问。某些 RDBMS 可用 GRANT 或 REVOKE 控制对表单个列的访问。

5. 数据定义语言(DDL)

其语句包括动词 CREATE 和 DROP。在数据库中创建新表或删除表(CREAT TABLE 或 DROP TABLE);为表加入索引等。DDL 包括许多与人数据库目录中获得数据有关的保留字。它也是动作查询的一部分。

6. 指针控制语言(CCL)

它的语句，像 DECLARE CURSOR、FETCH INTO 和 UPDATE WHERE CURRENT 用于对一个或多个表单独行的操作。

数据查询是通过 SELECT 语句来完成的。SELECT 语句可以从数据库中按用户要求检索数据，并将查询结果以表格的形式返回。

6.1.2 大数据时代的数据搜索

人类已经到了离开信息无法生活的地步。按照达尔文生物进化论，人类的信息吸收、筛选和处理的能力应该也会进化。人们对信息的需求并不会退化，反而会更加饥渴。搜索引擎需要解决的问题，不再是帮助人们从海量信息里面找到结果。而是在海量结果里面找到唯一。快速找到准确的答案比找到更多的答案更重要。

1. 结构化数据对搜索的价值

结构化数据和网页数据相比，更能满足第一点：找准唯一答案。网页分析是靠文本匹配。结构化数据的分析即支持内容提供者的主动接入，也支持搜索引擎的个性化精准分析。这两种方式都会增加内容提供者或者搜索引擎的成本，但是付出带来的回报是用户快速得到准确的唯一的答案。

2. 大数据挖掘是搜索引擎的机会

经过多年的发展，搜索引擎在文本分析、关系发掘、图谱构造、用户语义理解等方面已有丰富的积累。这些技术是大数据挖掘依赖的基本技术。我们会叫它挖掘引擎。而将挖掘和传统搜索结合起来，通过挖掘响应用户主动的或者被动的搜索需求，或许也可以称为“推荐引擎”。

一般来说，搜索引擎提供非结构化文本的查询服务，数据库引擎提供结构化数据的查询服务。因此结构化应用和利用数据库实现的数据挖掘过程难以拓展到非结构化数据上。比如搜索引擎对一个公开站点进行索引后，如果试图利用结构化数据分析方法来对网站的注册用户行为进行分析，通常是不太可能的。比如 BBS、博客和微博的顶贴人分析，哪些是假冒的明星粉丝，哪些人是“托儿”，对于一些商业化公司是有用的，特别是广告公司。

目前缺乏有效的手段来进行跨越站点的综合分析，一般是针对特定网站进行设计分析 程序。如果能够用搜索引擎来提供结构化查询的方法，很多标准的结构化分析程序将

可以派上用场。

如果说大数据是金矿，拥有大数据的垂直网站、社交网站、APP、云应用提供商、物联网拥有者、政府组织和企业即是金矿矿山的老板。他们可以自己从金矿里面掘金。也可以将金矿卖给搜索引擎或者大数据挖掘公司来挖掘。搜索引擎为金矿买单的同时，必须将自己从加速信息流动的管道，转变为会淘金的人。

3. 互联网信息的特点

1）面向显示与面向数据

从信息交换的角度看，目前互联网上的信息大多以 HTML 文档形式存在，用户与服务器之间信息的传递主要依赖超文本传输协议（HTTP）。HTML 文档中的信息是面向显示的，用规范的 HTML 标记 tag 定义文档的元数据（如标题 Title 等），或定义文档的文本应如何显示。这些标记的理解工作交由浏览器，而信息的理解工作则由用户自己完成。

XML 是互联网上信息交换的新标准，它支持用户自定义文档标记，用有序的、嵌套的元素组织有一定结构的数据，是面向数据的，程序可读解这些标记并依据标记的语义处理数据。以 XML 文档为主体的互联网将成为新一代以数据为中心的互联网计算环境。

2）半结构化与非结构化

在互联网上，数据嵌在 HTML 文档的文本中，而数据的部分组织信息嵌在标记中。从文档标记的角度看，HTML 显示超链接的文档；从数据的角度看，HTML 文档所蕴含的数据也是半结构化的，这是因为：

- 数据没有严格的结构模式；
- 含有不同格式的数据（如文本、声音、图像等）；
- HTML 文本无法区分数据类型；
- 多个异质数据源中不同的站点给相同的信息起不同的名字（如“级别”与“等级”等）。

目前，有很多研究正围绕半结构化数据和半结构化文档（如 SGML 或 XML 文档）的存储、模式、查询、优化等展开。

3）不同形式数据源的数据

除了保存在 HTML 文档中的信息外，互联网上还有大量信息存储在文本文档、传统的关系或对象数据库中，这些不同形式的数据在互联网上需要通过集成并用 HTML 文档显示，以实现共享和交换。

如何有选择地从已有数据开始，生成供浏览的页面并建立站点是互联网站点管理要考虑的问题。

4）静态与动态

互联网站点上的信息是随时间动态变化的，信息内容的变化（增删改）需要及时地反映到互联网页面中。另一方面，站点的页面组织结构可能发生的改变（如页面的增加、删除和修改）也要及时反映到站点页面的目录层次结构中。

由于站点的信息量大，手工动态改动信息的工作量很大，Web 站点管理应提供合适的工具进行站点维护或重构。

5）界面友好

Web 站点的信息主要面向一般的非计算机专业用户浏览和查询，因此，对界面的友好性、易用性提出了更高的要求。用户获取信息的渠道越来越多，方式越来越灵活，因此，提供给用户的服务应该适用于多种形式的用户界面。目前，很多搜索引擎通过 Form 的形式由用户填写搜索要求，这种用户界面虽然比较易用，但由于引擎搜索方式和搜索能力的限制，返回的结果形式单一、内容重复，并且没有智能化分析的功能，不能很好地满足用户的搜索要求。

4. XML 成为数据组织和交换事实上的标准

由于 Internet 的发展，网上数据不断激增，对网上信息的应用需求也不断提高，原有的对文本文件的链接浏览和关键词检索已无法满足一些复杂的应用需求。近年来，大量的研究致力于将数据库技术应用于网上数据的管理和查询，使查询可以在更细的粒度上进行，并集成多个数据源的数据。但是，将传统数据库技术直接应用于网上数据的最大困难在于：网上数据缺乏统一的、固定的模式，数据往往是不规则且经常变动的。因此，半结构化数据模型应运而生，其无模式及自描述的特点适用于描述网上数据。

事实上，日益普及的 XML 数据就是一种自描述的半结构化数据，它的出现推动了互联网在电子商务、电子数据交换和电子图书馆等多方面的应用。但对于如何有效地存储管理和查询这类数据，目前却莫衷一是，已有的数据库技术，如关系数据库、面向对象数据库，都不能完全适应于新的应用需求，而专用的半结构化数据管理系统目前仍处于初步实验阶段。

可以预言，XML 将成为数据组织和交换事实上的标准，大量的 XML 数据将很快出现在 Web 上。实质上，XML 为 Web 的数据管理提供了新的数据模型，很多成熟的数据库技术将进入 Web 信息处理领域，将其变为一个巨大的数据库。XML 是朝这个方向迈出的第一步。这种变化给数据库研究界带来了巨大的机会，使将数据库技术和研究扩展到对 Web 数据的管理成为可能。目前，对 XML 数据的存储和查询的研究方兴未艾。XML 数据模型与半结构化数据模型有着很多的相似性，可以说，XML 是互联网上的半结构化数据，它既为半结构化数据的研究展示了广阔的应用前景，同时也推动了半结构数据研究的发展。

6.1.3 数据库与信息检索技术的比较

互联网目前还只是一个巨大的分布的信息检索系统，大多数搜索引擎基于信息检索技术。数据库技术与信息检索技术有很多不同，详见表 6.1。

二者最重要的一个区别是数据库的数据结构性更强，比信息检索的数据包含更多的语义。在一定意义上，信息检索技术更适合于处理无结构数据，数据库则是管理结构数据的最好途径。在本质上，信息检索使用近似方法为用户的浏览需求查找相关信息。其中“近似”的含义包括近似的查询条件说明、近似匹配、近似结果。

表 6.1 数据库技术与信息检索技术比较

比较项目	数据库	信息检索(IR)
数据	有结构	无结构
模型	有确定性的模型	基于概率
查询语言	人工的(如 SQL 等)	自然的
查询规范	完全的	不完全的
匹配	精确匹配	部分匹配、最佳匹配
所需条目	基于匹配	基于相关
出错报告	敏感的	不敏感
推理	演绎	归纳
类属	单向度(Monothetic)	多向度(Polythetic)
数据更新	完全支持	不支持
事务	支持	不支持
使用	面向应用	面向人

数据库中简单演绎推理的形式为：如果 aRb 并且 bRc，那么 aRc。在信息检索技术中则经常使用归纳推理，关系只由确定或不确定的程度表达，因此，推理的可信度是个变量。这个区别导致数据库被描述为确定性的，而信息检索是概率性的。在信息检索中，经常用贝叶斯定理进行推导。

另外一个区别以类属为依据。数据库类属关系中的类由组成一个类的所有必要和充分的处理属性定义；在信息检索中，类的一个个体将只拥有该类所有个体的所有属性的一部分，类属没有充分或必要的属性。

数据库的查询语言通常是人工语言，有严格的语法和词汇表；在信息检索中，经常使用的是自然语言。

随着电子数据数量的激增和 Web 规模的快速增长，使用传统的信息检索方法在这样一个无限的信息海洋中要准确、快速定位所需信息时，越来越显得力不从心，在未来的 Web 发展中，如何提高信息检索的准确性和效率成为关键问题。另一方面，目前出现了超越浏览方式而使信息面向应用访问的迫切需求，从而为各种服务提供自主性、互操作性和 Web 意识。无结构的 HTML 文档及其相应的信息检索技术将不再适应下一代更复杂的 Web 应用。

因此，未来的 Web 信息将由更近似于数据库的方式进行管理，而不是目前采用的单一的信息检索方式。Web 资源需要以有结构的方式进行组织和访问。

6.1.4 数据库技术面临的 Web 数据管理问题

Web 目前的状况离 Web 上有效信息服务与信息管理的实现还有差距，这正为数据库技术向 Web 领域发展提供了空间。新环境中的数据库技术研究内容包括半结构化数

据模型及其理论、数据缓存与复制、事务管理、数据安全等，它与 Web 上已有的成熟技术（如信息检索技术）相结合，可以用来解决 Web 上数据管理、动态维护等关键问题。

1. 半结构查询语言与模式抽取

半结构化数据的研究起源于异质的数据源之间数据交换和集成，另外，一些数据源（如 Web）的数据并非像传统的结构化数据（如关系数据）那样有严格的数据格式和数据类型。半结构化数据的特点是没有事先给定的数据模式，或者数据模式对数据的约束不强，模式的规模比较大（有时甚至可以大过数据），模式是经常变动的，数据未赋予严格的类型。很多研究者研究了半结构化数据的存储、模式抽取、查询和用户界面等问题，并出现一些半结构化数据的原型系统，如 Lore。Lorel、UnQL 是比较典型的半结构化数据查询语言。

对结构化文档（如 SGML、XML 或 HTML）查询的研究，更多地考虑了对链接路径的查询能力、文字检索和字符串匹配能力，并考虑了结果的重构能力。

2. Web 站点建设与重构

Web 站点建设是从已有数据开始，创建用户可浏览的 Web 站点和 Web 站点视图。Web 站点重构是在已有站点的基础上，基于 Web 动态变化和安全的考虑，重构站点或 Web 站点的不同视图。Web 站点建设与重构既包括前面讨论的两个方面的问题，还包含其他方面的技术（如网络实现等），从数据的角度看，Web 创建者应考虑的问题有：

- 选择用于站点显示的数据。
- 确定 Web 站点的结构（页面的内容和页面之间的链接）和约束。
- 确定页面如何显示给用户。
- 信息集成技术是 Web 站点建设的基础，描述性的 Web 查询语言可以成为用于 Web 站点重构的方便和功能强大的工具。

3. 半结构化数据的存储研究

数据的存储研究包括两个问题：半结构化数据或 XML 数据的存储以及索引的存储。数据的存储有以下方式。

1）文本文件

文本文件是最简单、最直接地存储 XML 数据的方式。它与数据被理解的方式一致，自然地反映了对象之间的嵌套关系，且同一个对象的数据集中存储。缺点是存储粒度大，当数据量大时不利于实现网络通信和数据共享。

2）关系数据库

关系数据库存储半结构化数据或 XML 数据。可以利用数据库现有的存储管理、并发控制、恢复、版本机制等技术有效地管理数据。该方式的欠缺是一个简单的查询路径可能要通过多重链接实现，影响了查询的效率。半结构化数据的缺乏模式和数据类型的特性也使关系数据库的一些优化存储策略（如聚集存储等）不能应用。

3）面向对象数据库

很多商业的 XML 服务器采用这种方式。它利用 DTD 给出的类型信息构造类层次结构，正则表达式的符号可由基于对象数据模型的类型表达（如用 list 数据类型表达），也

可以通过创建新类实现(如“|”符号可用 union 类型的类实现)。该方式的数据模型更接近半结构化数据模型,并能更好地处理嵌套的集合和顺序,因此,其数据存储和查询处理可以用来提高 XML 或半结构化数据处理的效率。

4. 分布计算的研究

在信息分布的环境中,特别是在 Web 中,可能有两种情况出现:

1) 事先已知模式信息

知道数据如何分布,则可利用已知信息采取类似于分布关系数据库的半链接或半链接规约的技术进行查询处理。

2) 模式信息事先未知

需采用新技术处理。这时处理某一查询路径表达式比较好的解决办法是在每一个参与站点上建立一个对应于该路径表达式的自动机,各自将计算结果传到中心站点,然后计算出最终查询结果。这种方法可以减少不同参与站点间的通信次数。

5. Web 异构数据集成

Web 信息集成系统的目标是支持对 Web 上多个数据源的查询。它除与异构数据库集成系统相同外,还要处理大量的、数目递增的 Web 数据源,描述 Web 数据源特征的元数据很少,各数据源有很强的自治性。

建设 Web 信息集成系统有两种方法:数据仓库方法和虚拟方法。前者是将各数据源的数据装载到数据仓库中,用户的查询基于数据仓库的数据;后一种方法基于一个“中间模式”(Mediated Schema),数据仍保存在局部数据源中,通过各数据源的“包装程序”(Wrappers)将数据虚拟成中间模式,用户的查询基于中间模式,不必知道每个专门的数据源的特点,查询执行引擎直接与 Wrappers 打交道,将基于中间模式的查询转换为基于各局部数据源的模式。虚拟方法更适应于数据源数目多、各局部数据源的自治性很高且局部数据经常变化的 Web 环境。

6. Web 应用系统体系结构

Web 是一个分布的异质的计算环境,与这一环境相适应,其应用系统具有多层体系结构,即在客户/服务器两层结构之间具有若干个中间层。中间层的作用是集成、转换多个数据源的数据。中间层有两种实现方式。

1) 数据仓库

各数据源的数据被导入数据仓库中,实现数据集成并支持产生式系统的决策支持查询。这样的系统适合规模不很大但要求查询效率高,且源数据更新不多的情况。关键技术是有效的数据加载和增量更新维护。

2) 中介(Mediator)系统

数据并不实际存储在中间层,客户端发来的查询由中介系统转换为各数据源的查询。这种方法可适用于规模很大但对查询效率要求不高并且源数据经常更新的系统。关键技术是查询重写。

6.2 网络数据索引与查询技术

6.2.1 搜索引擎技术概述

网络数据查询目前使用的最多的是搜索引擎。搜索引擎(search engine)是指根据一定的策略、运用特定的计算机程序搜集互联网上的信息,在对信息进行组织和处理后,并将处理后的信息显示给用户,是为用户提供检索服务的系统。

1. 搜索引擎的发展

1990 年,加拿大麦吉尔大学(University of McGill)计算机学院的师生想到了开发一个可以用文件名查找文件的系统,开发出 Archie。当时,万维网(World Wide Web)还没有出现,人们通过 FTP 来共享交流资源。Archie 能定期搜集并分析 FTP 服务器上的文件名信息,提供查找分别在各个 FTP 主机中的文件。用户必须输入精确的文件名进行搜索,Archie 告诉用户哪个 FTP 服务器能下载该文件。

虽然 Archie 搜集的信息资源不是网页(HTML 文件),但和搜索引擎的基本工作方式是一样的:自动搜集信息资源、建立索引、提供检索服务。所以,Archie 被公认为现代搜索引擎的鼻祖。由于 Archie 深受欢迎,受其启发,1993 年又开发了一个 Gopher 搜索工具。

2. 搜索引擎分类

1) 全文索引

全文搜索引擎是名副其实的搜索引擎,国外代表有 Google,国内则有著名的百度搜索。它们从互联网提取各个网站的信息,建立起数据库,并能检索与用户查询条件相匹配的记录,按一定的排列顺序返回结果。

根据搜索结果来源的不同,全文搜索引擎可分为两类,一类拥有自己的检索程序(Indexer),俗称"爬虫"(Spider)程序或"机器人"(Robot)程序,能自建网页数据库,搜索结果直接从自身的数据库中调用,上面提到的 Google 和百度就属于此类;另一类则是租用其他搜索引擎的数据库,并按自定的格式排列搜索结果,如 Lycos 搜索引擎。

2) 目录索引

目录索引虽然有搜索功能,但严格意义上不能称为真正的搜索引擎,只是按目录分类的网站链接列表而已。用户完全可以按照分类目录找到所需要的信息,不依靠关键词(Keywords)进行查询。目录索引中最具代表性的有 Yahoo、新浪分类目录搜索。

3) 元搜索引擎

元搜索引擎(META Search Engine)接受用户查询请求后,同时在多个搜索引擎上搜索,并将结果返回给用户。著名的元搜索引擎有 InfoSpace、Dogpile、Vivisimo 等。

6.2.2 Web 搜索引擎工作原理

Web 搜索引擎的原理通常为:首先是用爬虫(Spider)进行全网搜索,自动抓取网页;

然后将抓取的网页进行索引,同时也会记录与检索有关的属性,中文搜索引擎中还需要首先对中文进行分词;最后,接受用户查询请求,检索索引文件并按照各种参数进行复杂的计算,产生结果并返回给用户。基于上面的原理,下面将简要介绍 Web 搜索引擎实现。

1. Web 搜索引擎的组成

搜索引擎一般由搜索器、索引器、检索器和用户接口四个部分组成,如图 6.1 所示。

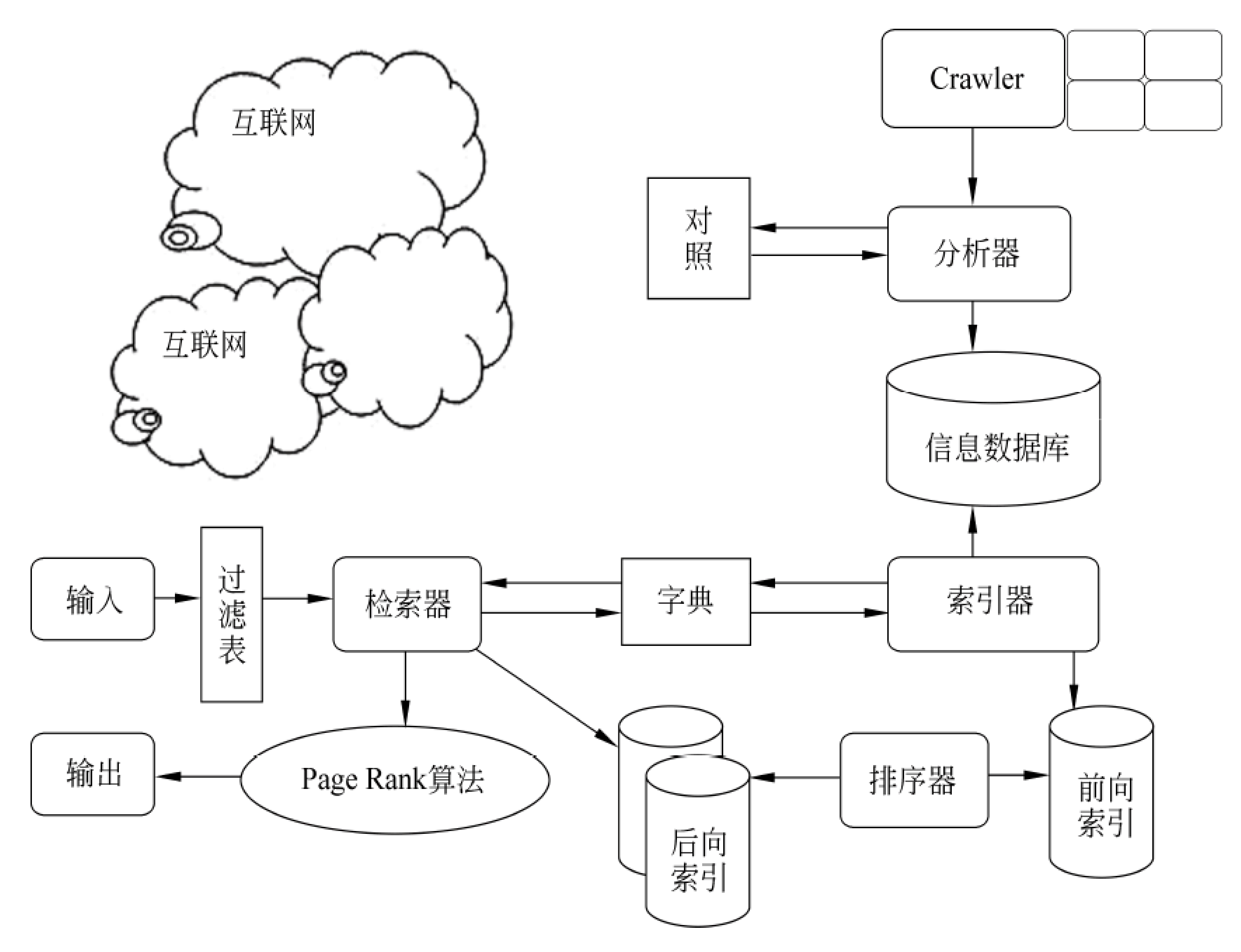


图 6.1　搜索引擎组成

(1) 搜索器:其功能是在互联网中漫游,发现和搜集信息;

(2) 索引器:其功能是理解搜索器所搜索到的信息,从中抽取出索引项,用于表示文档以及生成文档库的索引表;

(3) 检索器:其功能是根据用户的查询在索引库中快速检索文档,进行相关度评价,对将要输出的结果排序,并能按用户的查询需求合理反馈信息;

(4) 用户接口:其作用是接纳用户查询、显示查询结果、提供个性化查询项。

2. Web 搜索引擎的工作模式

(1) 利用网络爬虫获取网络资源。

这是一种半自动化的资源(由于此时尚未对资源进行分析和理解,不能成为信息而仅是资源)获取方式。所谓半自动化,是指搜索器需要人工指定起始网络资源 URL (Uniform Resource Locator),然后获取该 URL 所指向的网络资源,并分析该资源所指向的其他资源并获取。

网络爬虫访问资源的过程,是对互联网上信息遍历的过程。在实际的爬虫程序中,为了保证信息收集的全面性、及时性,还有多个爬虫程序的分工和合作问题,往往有复杂的控制机制。如 Google 在利用爬虫程序获取网络资源时,是由一个任务管理程序负责任务

的分配和结果的处理，多个分布式的爬虫程序从管理程序活动任务，然后将获取的资源作为结果返回，并从新获得任务。

其基本流程如图 6.2 所示。

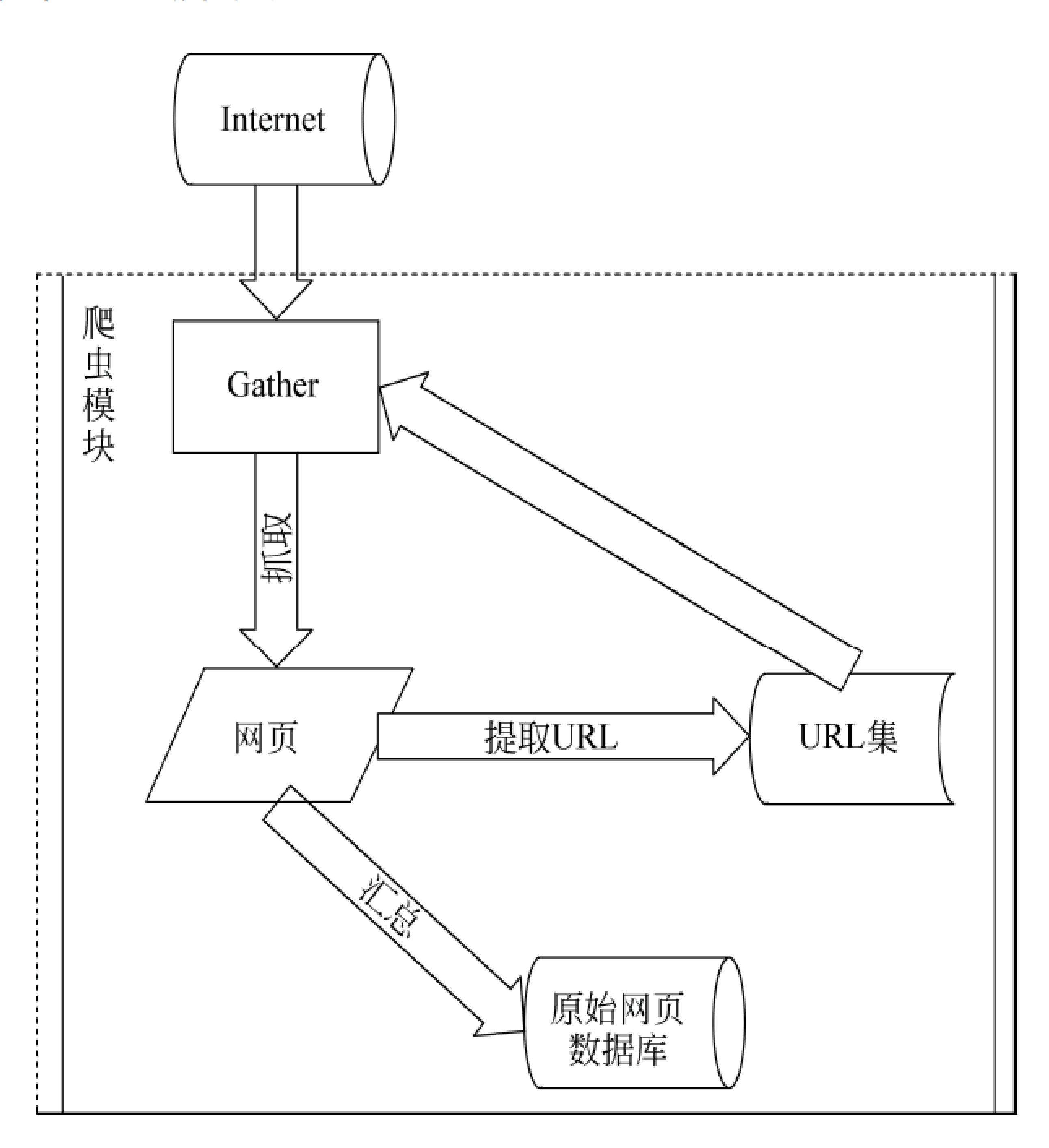


图 6.2 基本搜索器流程图

(2) 利用索引器从搜索器获取的资源中抽取信息，并建立利于检索的索引表。

当用网络爬虫获取资源后后，需要对这些进行加工过滤，去掉网控制代码及无用信息，提取出有用的信息，并把信息用一定的模型表示，使查询结果更为准确。其中信息的表示模型一般有布尔模型、向量模型、概率模型和神经网络模型等。

Web 上的信息一般表现为网页，对每个网页，应生成一个摘要，此摘要将显示在查询结果的页面中，告诉查询用户各网页的内容概要。模型化的信息将存放在临时数据库中，由于 Web 数据的数据量极为庞大，为了提高检索效率，须按照一定规则建立索引。

不同搜索引擎在建立索引时会考虑不同的选项，如是否建立全文索引、是否过滤无用词汇、是否使用 meta 信息等。

索引的建立包括：

- 分析过程，处理文档中可能的错误；
- 文档索引，完成分析的文档被编码进存储桶，有些搜索引擎还会使用并行索引；
- 排序，将存储桶按照一定的规则排序；
- 生产全文存储桶。最终形成的索引一般按照倒排文件的格式存放。

(3) 检索及用户交互。

前面两部分属于搜索引擎的后台支持。本部分在前面信息索引库的基础上，接受用户查询请求，并到索引库检索相关内容，返回给用户。这部分的主要内容包括：

用户查询(query)理解，即最大可能贴近地理解用户通过查询串想要表达的查询目

的，并将用户查询转换化为后台检索使用的信息模型；

根据用户查询的检索模型，在索引库中检索出结果集；

结果排序：通过特定的排序算法，对检索结果集进行排序。

现在用的排序因素一般涉及查询相关度，如 Google 发明的 pagerank 技术、百度的竞价技术等。由于 Web 数据的海量性和用户初始查询的模糊性，检索结果集一般很大，而用户一边不会有足够的耐性逐个查看所有的结果，所以怎样设计结果集的排序算法，把用户感兴趣的结果排在前面就十分重要。

Web 搜索引擎的工作模式如图 6.3 所示。

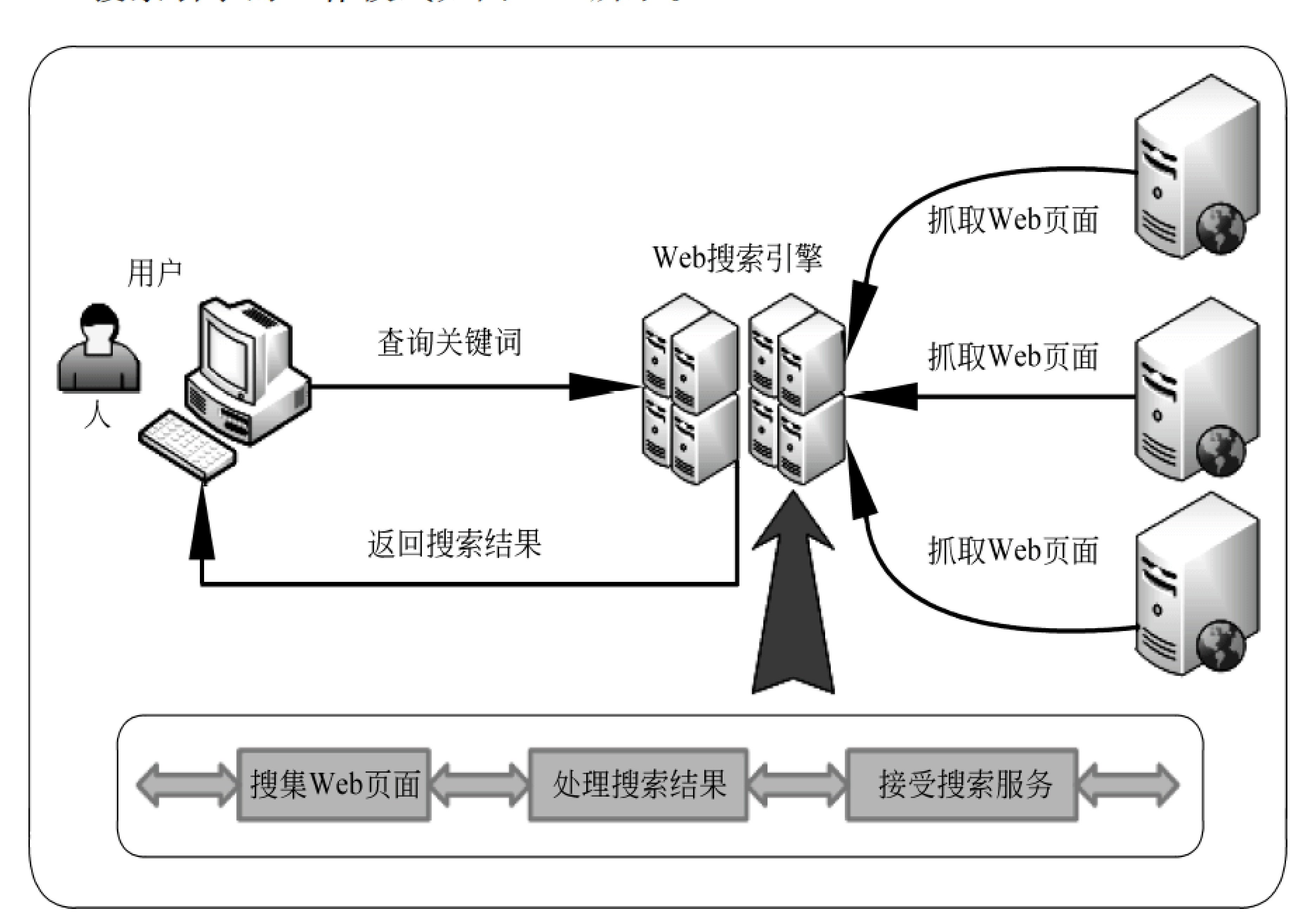


图 6.3　Web 搜索引擎的工作模式

3. 搜索引擎的技术设计与算法

搜索引擎的评价指标有响应时间、查全率、查准率和用户满意度等。其中响应时间是从用户提交查询请求到搜索引擎给出查询结果的时间间隔，响应时间必须在用户可以接受的范围之内。查全率是指查询结果集信息的完备性。查准率是指查询结果集中符合用户要求的数目与结果总数之比。用户满意度是一个难以量化的概念，除了搜索引擎本身的服务质量外，它还和用户群体、网络环境有关系。在搜索引擎可以控制的范围内，其核心是搜索结果的排序，即前面提到的如何把最合适的结果排到前面。

总的来说，Web 搜索引擎的三个重要问题是：

- 响应时间——一般来说合理的响应时间在秒这个数量级。
- 关键词搜索——得到合理的匹配结果。
- 搜索结果排序——如何对海量的结果数据排序。

所以搜索引擎的体系结构的设计时需要考虑信息采集、索引技术和搜索服务三个模块的设计。

1）信息采集

Web 搜索引擎的信息采集模块的主要功能是：

执行基于超文本传输协议(Hypertext Transfer Protocol,HTTP),从 Web 上收集页面信息,即 Web 机器人(爬虫)程序。

典型的基于超文本传输协议的网络应答图示见图 6.4。

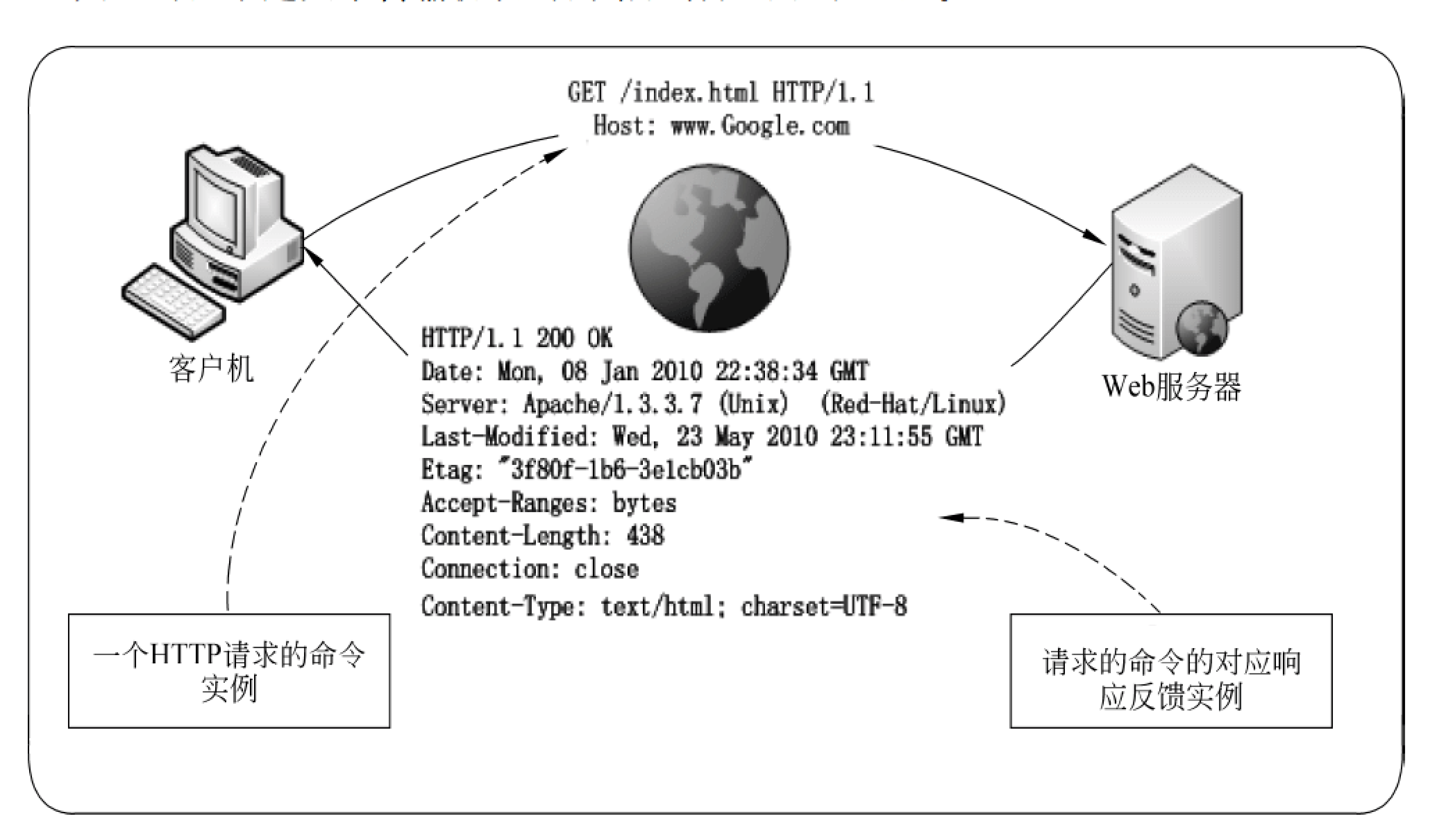


图 6.4 基于超文本传输协议的网络应答图

2）索引技术

(1）网络爬虫程序的工作模式。

网络爬虫程序根据 HTTP 协议,发送请求,并通过 TCP 连接接收服务器的应答。

由于 Web 搜索引擎需要抓取数以亿计的页面,所以建立快速分布式的网络爬虫程序才能满足搜索引擎对性能和服务的要求,其物理实现可能是一组终端。

爬虫程序的物理设备架构图如图 6.5 所示。

(2）网络爬虫程序的基础结构。

首先网络爬虫程序从 URL 链接库读取一个或多个 URL 作为初始输入并进行域名解析。

然后根据域名解析结果(IP)访问 Web 服务器,建立 TCP 连接,发送请求,接收应答,存储接收数据,并分析提取链接信息(URL)放入 URL 链接库里。

爬虫程序递归执行该过程直到 URL 链接库为空。网络爬虫程序的基础结构如图 6.6 所示。

3）信息采集优化

信息采集优化需要考虑到：网络连接优化策略、持久性连接和多进程并发设计等方面的问题。同时由于网络爬虫程序会频繁调用域名系统,域名系统缓存可提高爬虫程序性能需要使用 Web 缓存技术,如相关域名系统的缓存策略。

- LRU(Least Recently Used)算法：将最近最少使用的内容替换出 Cache 缓存；

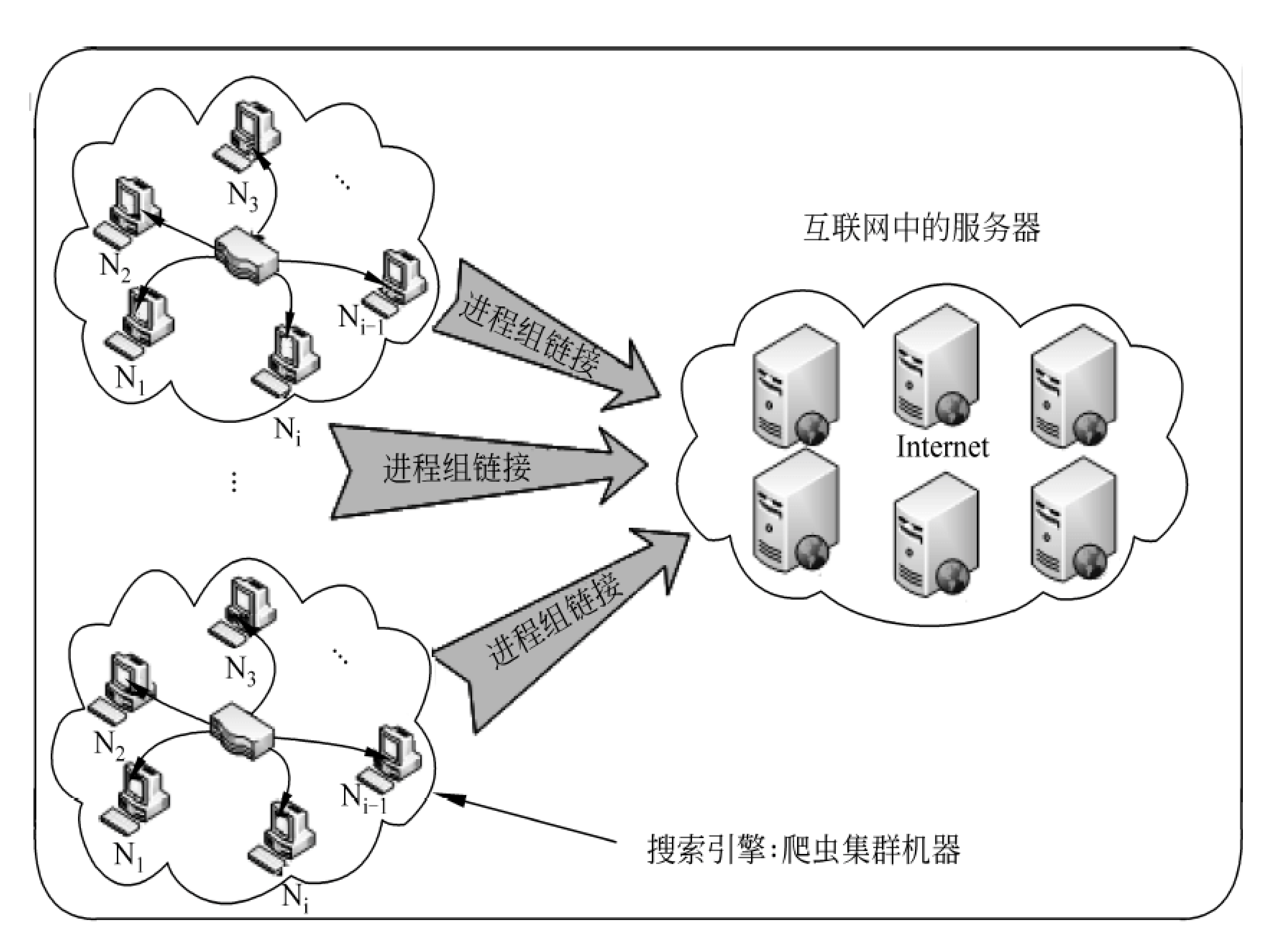


图 6.5　爬虫程序物理设备架构图

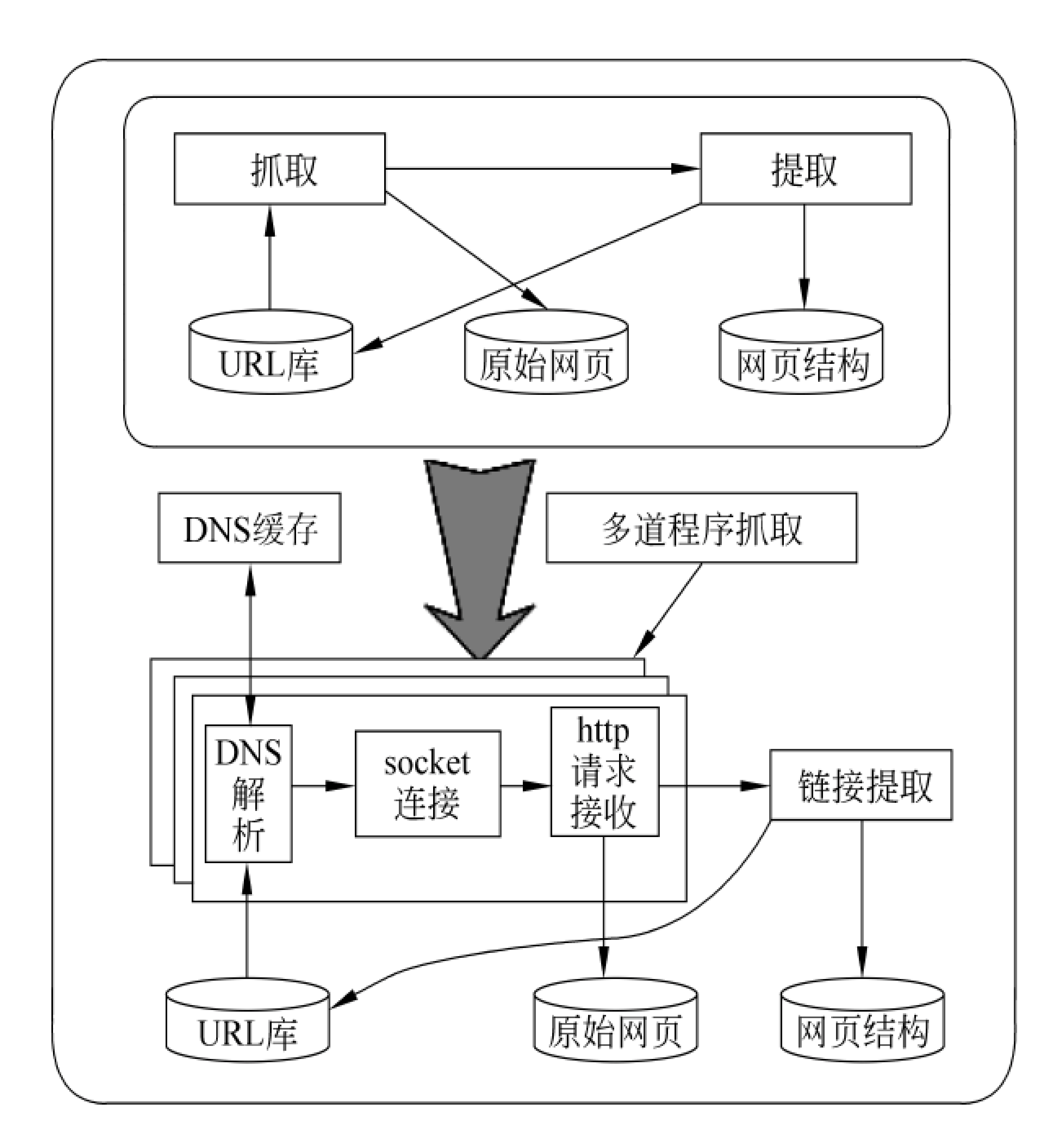


图 6.6　网络爬虫程序的基础结构

- LFU(Lease Frequently Used)算法：将访问次数最少的内容替换出 Cache 缓存；
- FIFO(First-In,First-Out)算法：在 Cache 缓存中执行数据的先进先出流程方法。

4）网页抓取算法

(1）深度优先算法。

在 Web 收集页面信息时，使用一个或一组预定义 URL 地址开始，然后根据页面内容

中的超链接深度抓取页面，直到搜索结束(没有新的 URL)。

(2) 广度优先算法。

在 Web 收集页面信息时，使用一个或一组预定义 URL 地址开始，然后根据页面内容中的超链接广度抓取页面，抓取下一层的 URL 直到这一层的 URL 完全被抓取，直到搜索结束时返回。

(3) 基于内容算法。

根据关键字、主题文档的相似度和链接文本(Linked texts)估计链接值，并确定相应搜索策略的算法。

链接文本是包含对 URL 链接解释说明和内容摘要的文字信息。

(4) 基于 HITS 的算法。

该算法的主要思想是：在抓取 Web 页面时，采用 Authority/Hub 抓取策略。Authority 表示该页面被其他页面所引用的次数(页面入度值，in-degree value)。Hub 表示其他页面引用该页面的次数(页面出度值，out-degree value)。

(5) PageRank(Google 的专利技术)。

Google 的 PageRank 根据网站的外部链接和内部链接的数量和质量来衡量网站的价值。PageRank 背后的概念是，每个到页面的链接都是对该页面的一次投票，被链接的越多，就意味着被其他网站投票越多。这个就是所谓的“链接流行度”——衡量多少人愿意将他们的网站和你的网站挂钩。PageRank 这个概念引自学术中一篇论文的被引述的频度——即被别人引述的次数越多，一般判断这篇论文的权威性就越高。

Google 有一套自动化方法来计算这些投票。Google 的 PageRank 分值从 0～10；PageRank 为 10 表示最佳，但非常少见，类似里氏震级(Richter scale)，PageRank 级别也不是线性的，而是按照一种指数刻度。这是一种奇特的数学术语，意思是 PageRank4 不是比 PageRank3 好一级——而可能会好 6～7 倍。因此，一个 PageRank5 的网页和 PageRank8 的网页之间的差距会比你可能认为的要大得多。

PageRank 的定义：

我们假设有 $T_1,\cdots,T_n$ 个页面指向页面 A(即引用)。参数 d 是一个阻尼因子，其取值区间属于(0,1)，我们通常取值为 0.85。$C(A)$定义为指向页面 A 的其他页面的连接数，页面 A 的 PageRank 或 PR(A)值可以通过下面的公式得到：

$$\mathrm{PR}(A) = (1-d) + d\left(\frac{\mathrm{PR}(T_1)}{C(T_1)} + \cdots + \frac{\mathrm{PR}(T_n)}{C(T_n)}\right)$$

注意：PageRank 值是 Web 页面的概率分布表示，所以所有 Web 页面的 PageRank 值的和是 1。

5) 索引技术

Web 爬虫抓取回来的页面信息，需要放入索引数据库里。索引建立的好坏对于搜索引擎有很大的影响，优秀的索引能够显著地提高搜索引擎系统运行的效率及检索结果的品质。文本分析技术是建立数据索引信息的支撑技术。

(1) 索引建立：预处理。

当 Web 搜索引擎获得数据信息以后，首先需要对数据进行预处理，如将句子切分成

有意义的词汇。由于中文的特殊性在切分句子时会产生二义性，如何合理地切分词汇是一个技术难题。

中文分词完全不同于英文分词，英文行文中，单词间以空格分隔；而中文只有字/句/段有明显的分隔符，唯独词没有形式上的分隔符存在。

(2) 索引建立：倒排文件模型。

(3) 倒排文件(inverted file)，是指一个词汇集合 W 和一个文档集合 D 之间对应关系的数据结构。建立倒排文件索引是建立索引数据库的核心工作。倒排文件模型如图 6.7 所示。

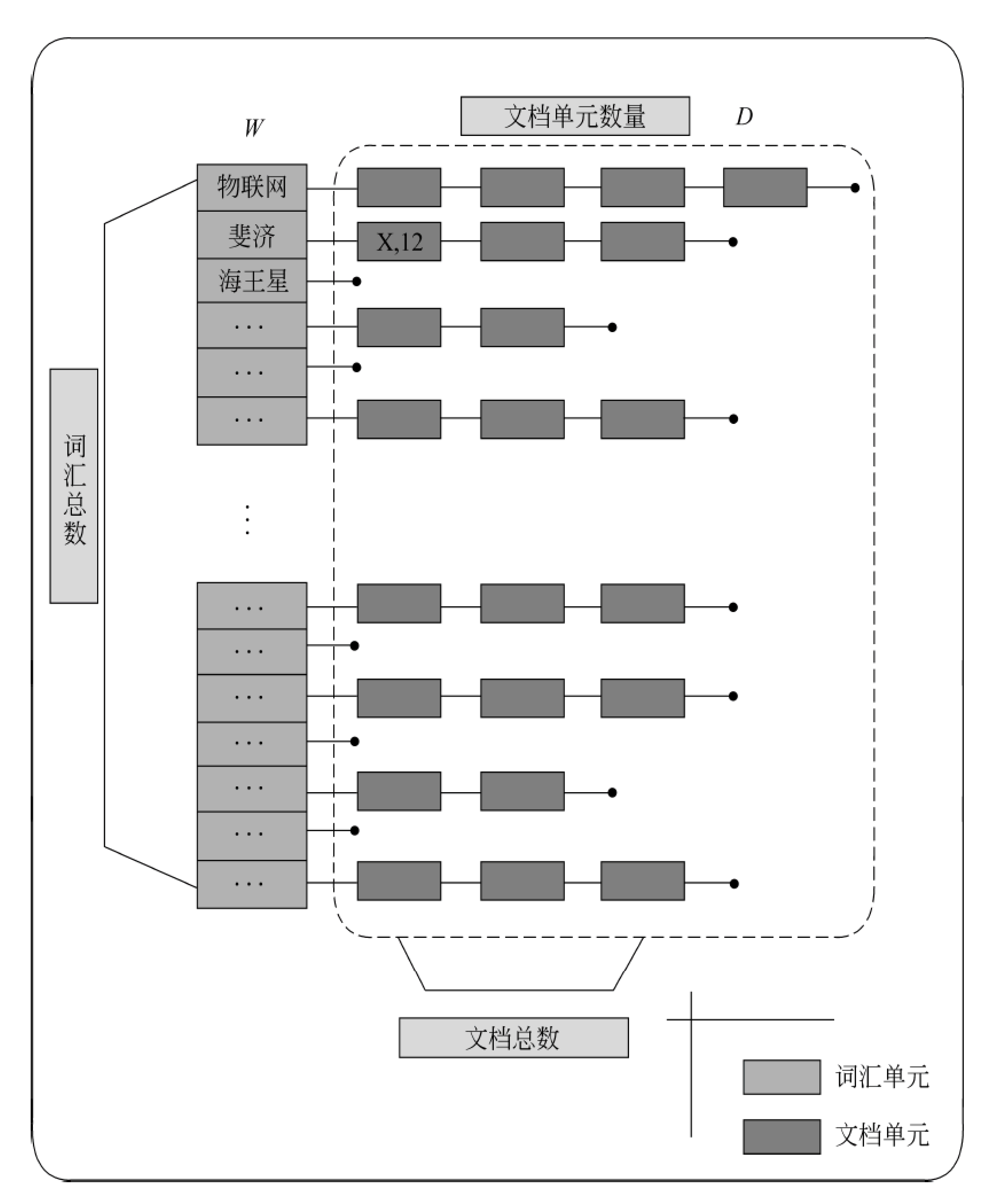


图 6.7　索引模块架构

6) 搜索服务

搜索服务是 Web 搜索引擎工作流程的最后一步，根据用户提交的查询关键字展开搜索，将匹配结果返回给用户。搜索服务的好坏直接影响 Web 搜索引擎的用户满意程度。

(1) 结果显示。

接受用户的输入，提交用户搜索请求。然后根据搜索结果列表合理的展示给用户。

并在保护隐私的前提下，记录用户使用行为的详细信息，以便提高下次服务的满意度。

(2) 网页快照。

Web 上的数据每时每刻都在变化着，所以随时存在着检索到的页面信息已经不存在的可能。Web 搜索引擎为了提高服务质量，需要对搜索到的页面信息进行快照，以便在原来页面信息失效的情况下，保证用户能够通过快照功能查看页面。

6.3 大数据索引与查询技术

6.3.1 大数据索引和查询

索引和查询技术是数据处理系统的重要入口之一，近年来随着数据量(Volume)、数据处理速度(Velocity)和数据多样性(Variety)的快速发展，大数据相关的索引和查询技术作为大数据的主要入口之一也变得更为重要。传统的索引和查询技术虽然不能很好地应对大数据带来的挑战，然而其核心技术，例如数据库和数据挖掘系统中使用的经典索引，例如哈希索引、B 树索引、位图索引和 R 树索引，信息检索系统中的倒排索引等依然是大数据索引和查询系统的基石。

大数据带来的主要挑战是其庞大的数据量，单个结点不能或者无法有效地处理这种数量级的数据。此外数据增长速度非常快，这要求系统不但能处理已有的大数据，还要能快速地处理新数据。这些特征使得我们需要考虑很多在大数据环境中独有的因素来开发和选择大数据索引和查询技术。

分布式是处理大数据的一个基本思路，这同样适用于大数据索引和查询系统。分布式索引把全部索引数据水平切分后存储到多个结点上，这可以很好地解决两个问题：

(1) 单个结点无法存储庞大的索引数据；

(2) 单个结点构建索引的效率瓶颈。当业务增长，需要索引更多的数据或者更快的索引数据时，可以通过水平扩展增加更多的结点来解决。

切分数据的方式有多种，常见的方法有随机方法、哈希方法和区间方法。随机方法将所有数据随机分布到不同的结点，这种方法不支持更新操作。哈希方法根据某个列或者某些列(称为分布键)的哈希值将数据分布到不同的结点。区间方法将所有的数据按照不同区间分布到不同的结点。区间到结点的映射信息需要保存下来。

不管使用什么样的切分方法，都需要注意数据分布的均匀性，避免大量数据分布到一个或者几个结点上，这样就失去了分布式计算的优势，因而对算法的选择和设计有一定要求。另外分布键的选择也很重要，好的分布键能将数据相对均匀地分布到不同的结点，从而达到负载均衡的目的。

由于索引数据是分布在不同的结点上，因而查询也是分布式的。所有结点或者部分结点的查询结果由主结点(主从架构)或者查询结点(点对点架构)进行汇总，然后得到最终结果。不同的分布式系统支持不同类型的查询语言和查询能力。分布式数据库系统支持 SQL 查询。

NoSQL 产品类型和功能各异，有的仅支持主键查询，有的支持范围查询，有的还支持

有限的 JOIN；全文检索系统的查询语法最为灵活，但通常不支持 JOIN 或者有限地支持 JOIN。

当一个结点故障时，将无法访问该结点上的数据。为了提高可用性，防止单点故障，通常使用镜像技术或者保存多个副本到不同的结点上。副本可以使用不同的分布策略，例如基于 Hadoop 的系统通常有两个副本：一个副本在同机架上的其他结点，另一个副本在其他机架的结点上。这样一方面可以有效利用数据局部性原理改进性能，另一方面可以最大化地保证数据的可用性。

有些系统副本仅仅起到数据备份的作用，这种类型的副本不能接受查询请求，主要目的是提高系统的可靠性。有的系统的副本还可以处理用户查询请求，从而实现负载均衡以最大化地利用系统资源。然而副本的引入也大大增加了系统的复杂性，因为分布式环境下任何一个结点可能在任何时刻出错：网络可能故障、磁盘可能故障、系统可能崩溃。

多数系统采取保证数据高度一致性的策略：只有主副本接受写请求，然后通过文件块复制或者写管道将数据写入到其他副本。也有一些 NoSQL 系统采用最终一致性策略，这种策略中在某一个时刻数据在不同的副本上可能是不一致的，但是当没有对该数据的更新时，最终的访问将返回该数据的最新值。

当系统不能适应业务的需求时，需要对系统进行动态扩容，这通常需要进行数据的再分布，即根据新系统中结点的个数按照数据分布策略重新对数据进行分布。当数据量庞大时，扩容可能需要较多的时间。为了降低需要移动的数据量，可以采取某些算法来实现，例如一致性哈希算法。

目前各大数据库厂商，例如 Oracle、IBM、Greenplum 都已经有支持分布式索引和查询的产品，很多 NoSQL 数据库例如 Mongodb、HBase、Cassandra 也支持分布式索引和查询。

还有很多面向全文检索的产品，例如 Solr、ElasticSearch、Sphinx 均支持分布式全文索引和查询，且这些产品都是开源的。值得一提的是，Greenplum 的 GPText 将 Solr 的全文检索能力引入到了 Greenplum 数据库之中，使得它可以同时支持 SQL 和 Solr 的全文检索。

6.3.2　大数据处理案例：登机牌、阅卷与 MapReduce

映射-归约(MapReduce)是 Google 多年前推出的建立海量数据索引的方法，有人说它是里程碑性的技术。而理解“映射-归约”，又是理解更时髦的 Hadoop 和 Spark 等大数据技术的基础。其实，在 Google 之前，人们就不知不觉地用了映射-归约技术，如机场分发登机牌、银行取号排队、流水作业阅卷。

1. 搜索引擎有多快

以下将三次用到飞机航班相关的实例，在百度(或 Google)查询栏中输入 CA1209，不到一秒钟，百度给出 200 个结果，分成 20 多页呈现，为后面叙述方便，不妨把这 200 个结果页面记为 p1，p2，…，p200，如图 6.8 所示。

2. 为什么快？养兵千日的倒排索引

搜索网站服务器中有这样一个索引，类似于规范的科技书籍之书末索引，其特点是一

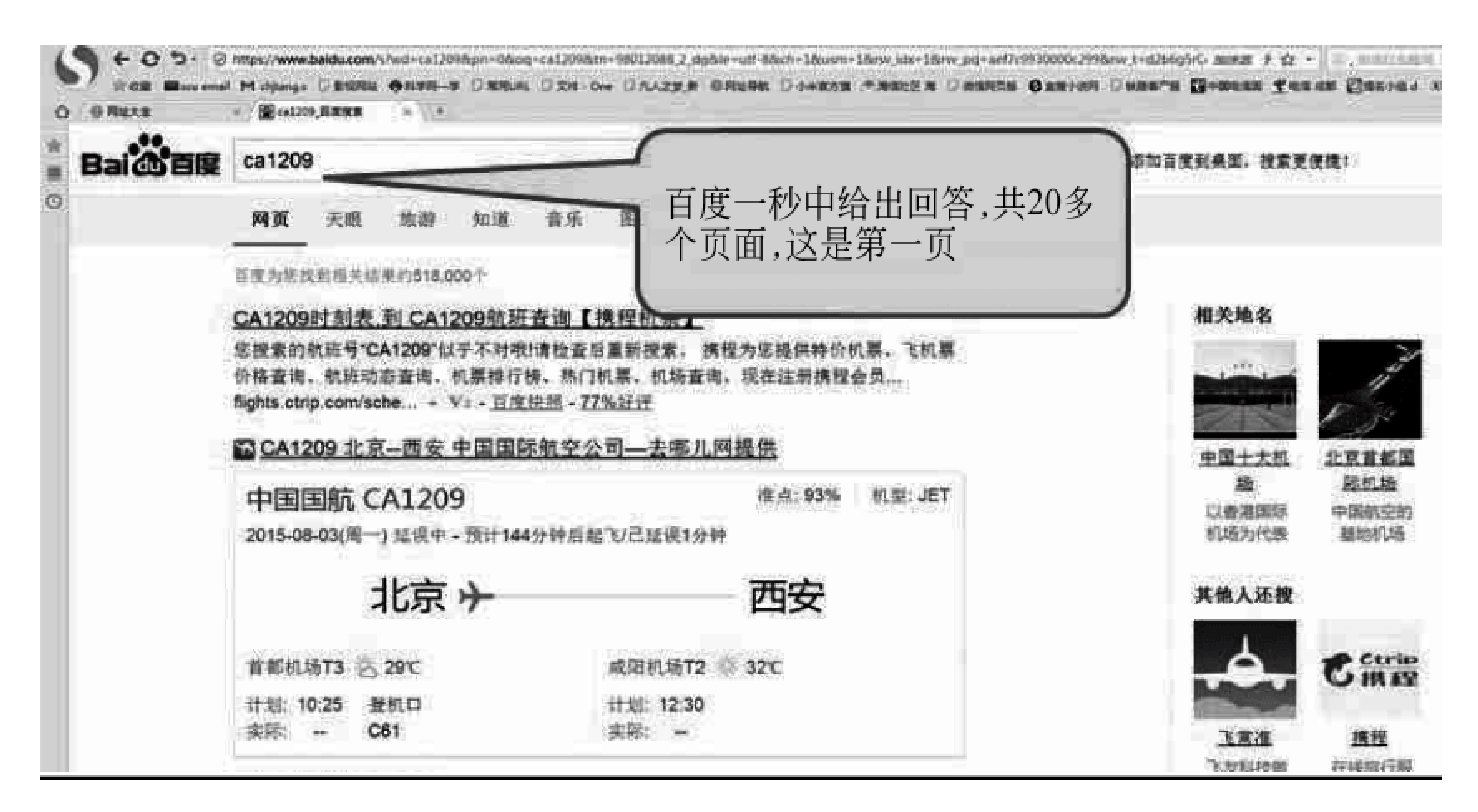


图 6.8 搜索引擎有多快

个关键字对多个标号(或页码),又称为倒排表,其中航班 CA1209 这一项关键字,对应了百度列出的 200 条信息 p1,p2,…,p200。

百度在回答查询时,一秒钟送出这些现成的 p1,p2,…,p200,如表 6.2 所示。

表 6.2 倒排索引

关　键　字	包含关键字的页面队列
…	
航班 CA1209	p1,p2,…,p200
…	…

而这个倒排索引是由若干万台计算机(或 CPU)以 365 天×24 小时方式,夜以继日得出的结果。

3. 大数据环境下,倒排索引有多难

设某搜索引擎每天新增 1 亿篇网文,考虑到网文中有些太平凡的字词(停用词,Stop Word)不适合做关键字,如"的""地""得""不但""而且",等等,每个网页平均有效关键字按 100 估算,要做完一天新增网页的倒排表,用笨方法,需要读扫描 1 亿网页,写处理 100 亿词汇,然后记录下所有如下的数据对:

<关键字,所在页面>

再加以整理、去重、合并、压缩,这需要用多少个 CPU 小时! 需要多大的空间!

Google 在创业之初,提出了一个从海量文档中做倒排索引的聪明方法——Map Reduce(映射-归约),正是它,协调若干万台计算机,并行计算,完成了倒排表的构建与维护,使 Google 在求多求快的竞争中立于不败之地。

下面用机场办理登机牌的例子来说明。

4. 机场登机牌分发中的映射-归约

乘客在首都机场办理登机手续时,会经过三次映射(三次映射的复合还是映射)和一

次归约。

(1) 第一次映射，分而治之，进入首都机场候机大厅，乘客会看到如图 6.9 所示的液晶屏：



图 6.9　机场登机牌分发中的映射-归约

这屏信息提示乘客按航班分流，例如航班 CA1209 是在 K0～K14 号的 15 个值机台办理登机牌；分而治之，缩小了数据规模，这是古代政治家治理国家的经典策略，也是如今处理大数据的朴素方法。

(2) 第二次映射，把乘客分到值机台。

图 6.10 展示了首都机场 K0～K14 值机台办理登机牌的情况。为保护隐私，故意把图片做了模糊化处理。



图 6.10　第二次映射

右边是乘客队列(相当于第 3 段例子中的每天新增的 1 亿个网页)。在中间，一位机场人员把乘客分成组(例如 15 人一组)，一次进入一组，分到 15 个值机柜台，引导加上乘客趋短避长的心态，保证了各个小队列长度大致平衡。

(3) 第三次映射,把乘客映射到《航班,座号》。

柜台处理包括验看证件,发放登机牌,把乘客分到航班上,并给托运行李挂上航班标签。

设在多个值机台的并行工作下,证件号为1、3、5的乘客,分到了航班CA1209,而证件号为2、4、6的乘客,分到了航班3U8882,于是,得到了下列《乘客,航班号,座号》三元组:

《1,CA1209,1排A》,《3,CA1209,2排B》,《3,CA1209,3排C》,

《2,3U8882,5排A》,《4,3U8882,7排B》,《6,3U8882,2排C》,

至此,并行地完成了这6位乘客的第三次映射。

(4) 归约成为倒排表。

把上述映射的结果按航班合并、约简,成为便于使用的倒排表,如表6.3所示。

表6.3 归约成为倒排表

关　键　字	乘客证件号及其座号队列
……	……
航班CA1209	1(1排A),3(2排B),5(3排C)……
航班3U8882	2(5排A),4(7排B),6(2排C)……
……	……

这一步骤,把同一航班的乘客归到一起,例如,1、3、5出现在倒排表中CA1208这一行右边,对乘客而言,是归类,对信息而言,是约简,把这一动作被称为归约(reduce),是再合适不过了。

登机牌在该航班起飞前半小时将停办,对应倒排表停止变化,把乘客按某指标(通常关注重要程度)排序,被分发到该航班和机场、保险公司等相关部门。

此外,用多个单关键字的倒排索引作交集,可以得到多关键字的倒排索引。

(5) 倒排表帮助改善服务上述倒排索引能帮助机组人员知道登机人数与座位,改善服务,例如,能叫出头等舱客户和金卡客户的姓名且服务到座位,就显得格外温馨和谐。

如有突发事件发生,作为"处突"依据,例如,马航官方能在突发事件后很快查出MH370的乘客信息。

综上所述,办理登机牌的全过程可以表达为如图6.11所示的经典MapReduce图,这个图大致反映了并行地映射-归约的流向,但未表达描述的归约细节。

现在的互联网搜索引擎,倒排表中的机理大致如上,但数量增大若干个数量级,相当于在图6.11中的乘客组有几千万,值机台(CPU)有100万,而航班(倒排索引项)是几万至几十万。

需要说明的是,这只是为了说明"映射-归约"机制而编的例子,真实的机场工作机制要复杂得多。

5. 安检时的映射-归约

在首都机场,可以看到,在安检时,还有一次MapReduce过程,源源不断的乘客乘坐

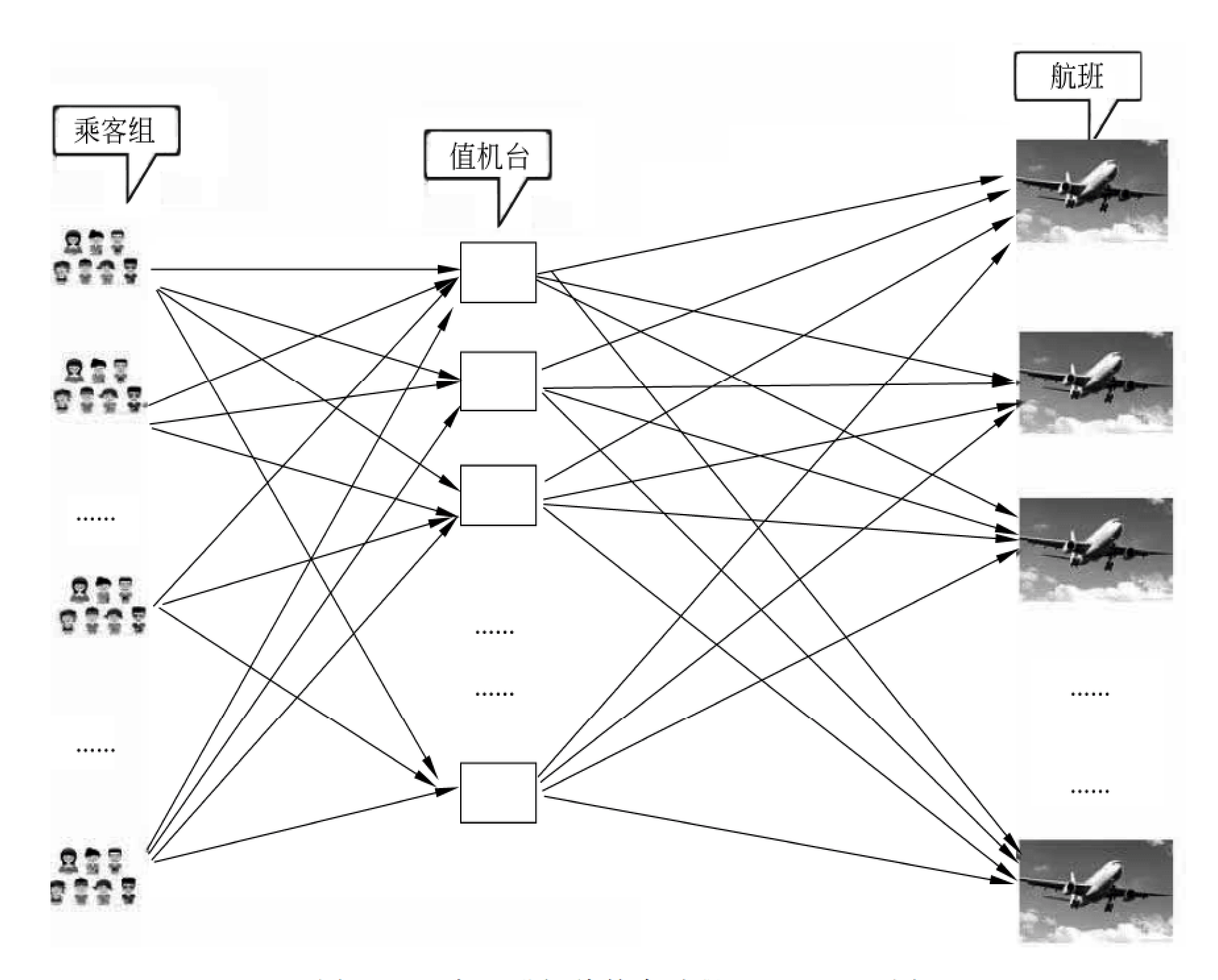


图 6.11 办理登机牌的全过程 MapReduce 图

扶梯下到安检大厅：

Map——一位安检人员指引乘客，分流到个安检口；

Reduce——安检后，分成若干类：大部分归约为 PASS 类，部分乘客有不合适行李，要做处理，或自弃，或托运，安检人员会对应机票、身份证做相应记录……

6. 映射-归约技术要点

上面的例子在思路上还真是 MapReduce（不仅仅是比喻），虽然还只是“小样”，但事不同而理同。

大数据中的映射-归约有下列要点。

(1) 目标：完成某一类计算，典型实例之一是生成某个关键字上的倒排索引；

(2) 对象：PB 级的数据，例如来自云、来自分布式文件系统的文档。

(3) 并行处理，多个（几百至几十万个，甚至更多）处理单元（计算机、CPU、人员）；

(4) 有序：在机场、车站，当客户增加，仅仅增加服务台来做归约（Reduce），常常不够有序，增加一个映射（Map）机制，把被处理对象分配到处理单元，是不可少的环节。春运中人们更体会到这一条。

(5) 多层映射，多层归约：在首都机场我们看到了映射有三层，第一次映射到值机台分区，分而治之；第二次次到值机台，第三次映射到《乘客，航班号，座号》三元组；根据实际情况，归约也可以是多层次的。

这里也要强调，小样和真实数据还有差距，量变超过了一定阈值，会引发质变，这一点在实践中必须注意。

6.4 相似性搜索工具

相似性搜索工具用于识别哪些候选要素与要匹配的一个或多个输入要素最相似(或最相异)。相似性基于数值属性(感兴趣属性)的指定列表。如果指定了一个以上的要匹配的输入要素,相似性将基于每个感兴趣属性的平均值。输出要素类(输出要素)将包含要匹配的输入要素以及找到的所有匹配的候选要素,这些要素以相似程度排序(由最相似或最不相似参数指定)。返回的匹配数基于结果数参数的值。

1. 可能的应用

可以用相似性搜索工具找出和某城市在人口、教育以及临近特定娱乐机会方面相似的其他城市。

当地领导干部可能希望促进其城市的潜在业务,从而提高税收。相似性搜索工具有助于帮助他们找出与其城市类似的城市,以便他们可以比较自身的吸引力属性(例如,低犯罪率和高成长率)。

这些领导干部也可能有兴趣查找比其城市大或小、但位置相似(余弦相似性)的城市。找出与他们的城市相似但更小或更大、并且具有他们期望拥有的商业吸引力的地方可以让他们指出相似性,同时可以强调小的优势(不那么拥堵、小城镇韵味)或者大的好处(例如更多的顾客)。

这些领导干部们还可能关注和他们的城市不特别相似的城市。如果任何不特别相似的地方表现出他们期望吸引的业务竞争优势,此分析则可以为他们提供相对所需的信息。

人力资源经理可能希望能够证明公司的工资范围。找出在大小、生活成本、市容建筑方面相似的城市后,便可以查看这些城市的工资范围,从而查看自己是否在此行列。

犯罪分析师希望搜索数据库以查看某罪行是否属于较重犯罪形式或有重罪趋势。执法机构用此方法揭露毒品种植地或生产地。标识具有相似特征的地方可能有助于制定未来的搜索目标。

大型零售商不仅拥有数个成功店铺,也有少数业绩不佳的店铺。找到一些具有相似人口特征和环境特征(交通便利性、知名度以及商业互补性等等)的地方有助于标识新店的最佳位置。

2. 匹配方法

匹配可基于属性值、等级属性值或属性剖面(余弦相似性)。下面介绍每种方法采用的算法。对于所有方法,如果有一个以上的要匹配的输入要素,则需要将这些要素的属性取平均值来创建复合目标要素,以用于匹配过程。复合目标要素如表 6.4 所示。

1) 属性值

为匹配方法参数选择 ATTRIBUTE_VALUES 时,工具首先标准化所有感兴趣属性。对于每个候选要素,将从目标要素中减去标准化值,求得平方差,然后再将每个平方差相加。相加的总和即为该候选要素的相似性指数。所有候选要素经处理后,按照指数从小(最相似)到大(最不相似)的顺序对候选要素进行分级。

表 6.4　复合目标要素

要匹配的输入要素	感兴趣属性		
	人口	工作人口	失业率
A	100 万	50 万	2.5%
B	105 万	40 万	2.6%
用于匹配的复合目标要素	102.5 万	45 万	2.55%

开始行动：

属性值的标准化涉及 Z 变换，即从所有属性值的平均值中减去每个属性值然后除以所有值的标准差。标准化将所有属性放在同一比例，即使它们由不同类型的数字表示时也是如此：比率（数字 0 到 1.0）、人口（数值大于 100 万）、距离（例如 1000m）。

2）等级属性值

为匹配方法参数选择 RANKED_ATTRIBUTE_VALUES 时，工具首先为目标要素和所有候选要素对感兴趣属性进行分级排序，然后为每个候选要素对目标要素相关的每个属性平方差求和。

3）属性剖面

为匹配方法参数选择 ATTRIBUTE_PROFILES 时，此工具首先将所有感兴趣属性标准化（此方法需要最少两个感兴趣属性）。然后用余弦相似性数学方法比较每个候选要素的标准化属性矢量与所匹配目标要素的标准化属性矢量。两个矢量 A 和 B 的余弦相似性按照如下方式计算：

$$\text{余弦相似性指数} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

余弦相似性与属性量的匹配无关，而此方法主要关注这些属性的关系。如果在比较的矢量（目标与候选要素之一）中创建标准化属性的剖面图（折线图），则可以看到非常相似或非常不同的剖面。创建标准化属性的剖面图如图 6.12 所示。

顶部一对属性的剖面非常相似，而底部一对属性的剖面十分不同。

余弦相似性指数范围为 1.0（完全相似）到 −1.0（完全不相似），并在 SIMINDEX（余弦相似性）字段中加以报告。可以使用此相似性方法以可能更大或更小的比例找出具有相同特征的地方。

3. 最佳做法

1）制图相似性模式

如果将结果数参数设定为 0，则工具将对所有候选要素进行分级排序。此分析的输出将显示相似性的空间模式。注意，在分级排序所有候选要素时，可以获取有关相似性和相异性的信息。显示相似性的空间模式如图 6.13 所示。

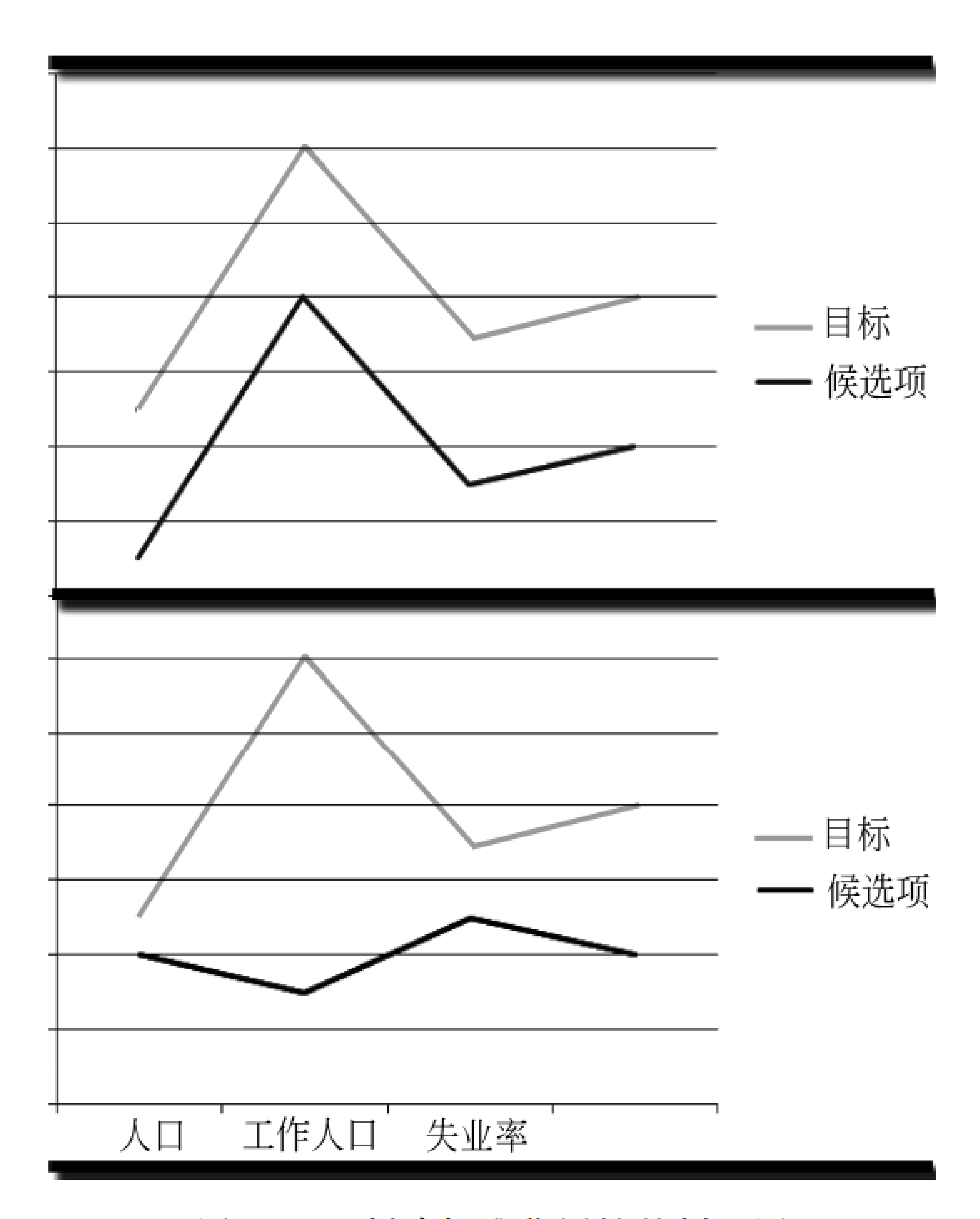


图 6.12　创建标准化属性的剖面图

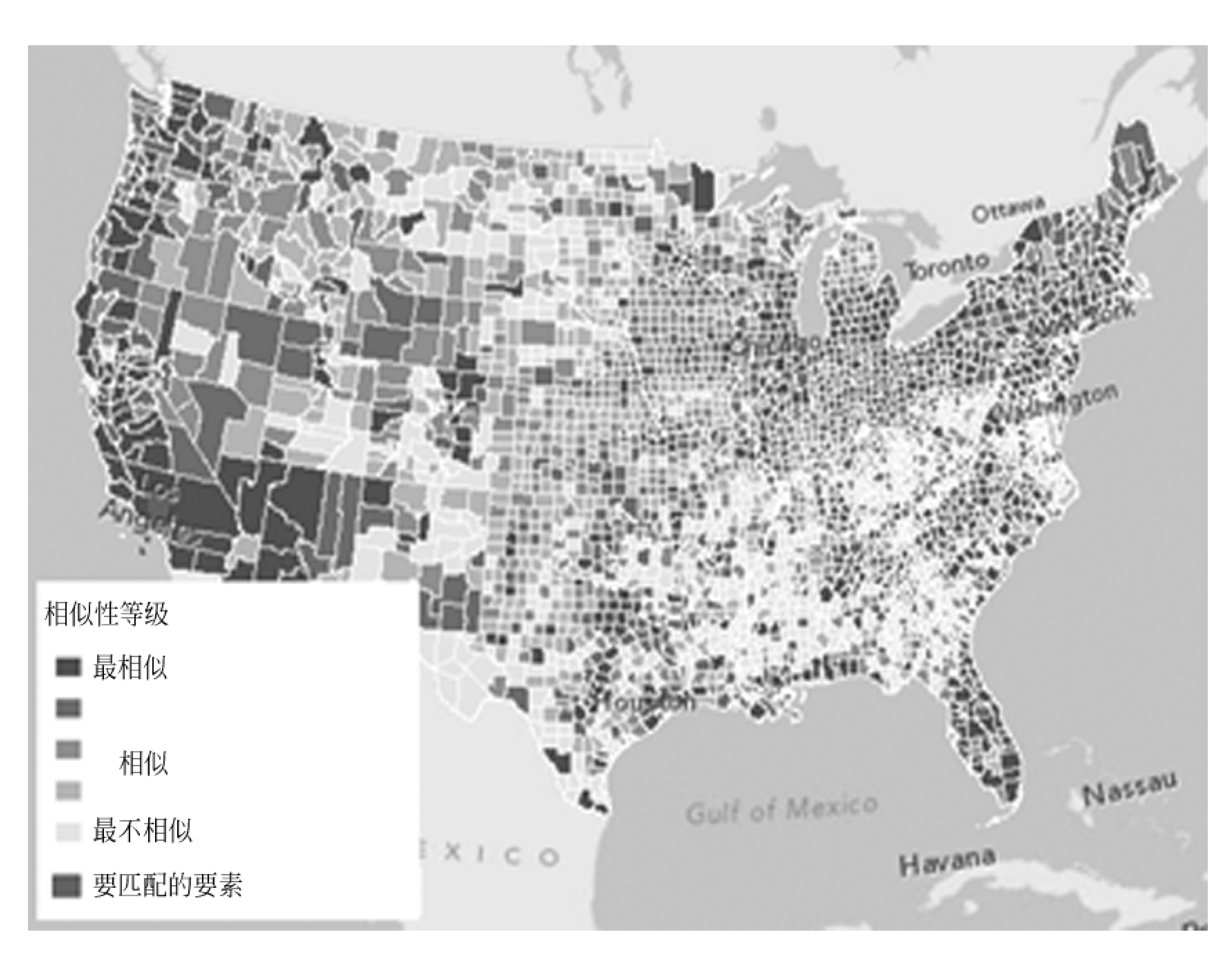


图 6.13　显示相似性的空间模式

2）包括空间变量

假设知道某濒危物种在某地(面区域)生存很好,希望找到该物种也可能茁壮成长的其他地方。你可能想寻找与物种成功存活环境相似的地方,但可能还需要这些地方足够

大，足够紧凑以保证物种成活。在此分析中，可以计算每个面区域的紧凑性指标（一般紧凑性测量基于与圆圈区域具有相同周长的面的面积）。运行相似性搜索工具时，可以将紧凑性测量和能够反应面的尺寸（Shape_Area）的属性包括在追加到输出的字段参数中。就紧凑性和面积排列出前 10 个匹配解决方案，将有助于识别再引入物种的最适宜位置。

或许你是一个对扩大业务感兴趣的零售商。如果你已经拥有成功店铺，可以通过能够反映成功关键特征的属性来帮助查找扩大业务的候选位置。假设你销售的产品对大学生最有吸引力，并且想避免靠近现有店铺或远离竞争者。在运行相似性搜索工具之前，可以使用近邻分析工具创建空间变量：与大学或大学生密度较大处之间的距离、与现有店铺的距离以及与竞争者的距离。运行相似性搜索工具时，可以将这些空间变量包括在追加到输出的字段参数之中。

6.5 数据展现与交互

计算结果需要以简单直观的方式展现出来，才能最终为用户所理解和使用，形成有效的统计、分析、预测及决策，应用到生产实践和企业运营中，因此大数据的展现技术，以及与数据的交互技术在大数据全局中也占据重要的位置。

Excel 形式的表格和图形化展示方式是人们熟知和使用已久的数据展示方式，也为日常的简单数据应用提供了极大的方便。华尔街的很多交易员还都依赖 Excel 和他们很多年积累和总结出来的公式来进行大宗的股票交易，而微软公司和一些创业者也看到市场潜力，在开发以 Excel 为展示和交互方式，结合 Hadoop 等技术的大数据处理平台。

人脑对图形的理解和处理速度，大大高于文字。因此，通过视觉化呈现数据，可以深入展现数据中的潜在的或复杂的模式和关系。随着大数据的兴起，也涌现了很多新型的数据展现和交互方式，和专注于这方面的一些创业公司。这些新型方式包括交互式图表，可以在网页上呈现，并支持交互，可以操作、控制图标、动画和演示。另外交互式地图应用，如 Google 地图，可以动态标记、生成路线、叠加全景航拍图等，由于其开放的 API 接口，可以与很多用户地图和基于位置的服务应用结合，因而获得了广泛的应用。Google Chart Tools 也给网站数据可视化提供了很多种灵活的方式。从简单的线图、Geo 图、gauges（测量仪），到复杂的树图，Google Chart Tools 提供了大量设计优良的图表工具。

诞生于斯坦福大学中的大数据创业公司 Tableau 正逐渐成为优秀的数据分析工具之一。Tableau 将数据运算与美观的图表完美地接合在一起。公司可以用它将大量数据拖放到数字“画布”上，转眼间就能创建好各种图表。Tableau 的设计与实现理念是：界面上的数据越容易操控，公司对自己在所在业务领域里的所作所为到底是正确还是错误，就能了解得越透彻。快速处理、便捷共享，是 Tableau 的另一大特性。仅需几秒钟，Tableau Server 就可以将交互控制面板发布在网上，用户只需要一个浏览器，就可以方便地过滤、选择数据并且对他们的问题得到回应，这将使得用户使用数据的积极性大大增加。

此外，3D 数字化渲染技术也被广泛地应用在很多领域，如数字城市、数字园区、模拟与仿真、设计制造等，具备很高的直观操作性。现代的增强现实 AR 技术，通过计算机技术，将虚拟的信息应用到真实世界，真实的环境和虚拟的物体实时地叠加到了同一个画面

或空间同时存在。结合虚拟3D的数字模型和真实生活中的场景,提供了更好的现场感和互动性。通过AR技术,用户可以和虚拟的物体进行交互,如试戴虚拟眼镜、试穿虚拟衣服、驾驶模拟飞行器等。在德国,工程技术人员在进行机械安装、维修、调式时,通过头盔显示器,可以将原来不能呈现的机器内部结构及其相关信息、数据完全呈现出来。

现代的体感技术,如微软的Kinect以及Leap公司的Leap Motion体感控制器,能够检测和感知到人体的动作及手势,进而将动作转化为对计算机及系统的控制,使人们摆脱了键盘、鼠标、遥控器等传统交互设备的束缚,直接用身体和手势来与计算机和数据交互。当今热门的可穿戴式技术,如Google眼镜,则有机地结合了大数据技术、增强现实以及体感技术。随着数据的完善和技术的成熟,我们可以实时地感知周围的现实环境,并且通过大数据搜索、计算,实现对周遭的建筑、商家、人群、物体的实时识别和数据获取,并叠加投射在人的视网膜上,这样可以实时地帮助我们工作、购物、休闲等,提供极大的便利。当然这种新型设备和技术的弊端也是显而易见,我们处在一个随时被监控、隐私被刺探、侵犯的状态,所以大数据技术所带来的安全性问题也不容忽视。

6.6 数据可视化

图灵奖得主Jim Gray在2007年提出了“以数据为基础的科学研究第四范式”的概念,研究方法已经从“我应该设计个什么样的实验来验证这个假设?”逐渐发展为“从这些已知的数据中我能够看到什么相关性?”数据可视化是获取大数据Value的有效手段。

6.6.1 数据可视化概念

1. 什么是数据可视化

数据可视化是关于图形或图形格式的数据展示。在一个被关注的连贯而简短的报告中体现大量的信息。虽然数据可视化可以处理书面信息,但焦点往往是使用图片和图像信息传达给观众。

此外,数据可视化不仅限于涉及数据的使用。也可能是可视化各种各样的信息——你可以将自己的想法与猜想与他人交流。如今,可以添加各种技术应用到数据可视化,甚至是选择交互式的可视化方法。

信息的视觉化表达是一种古老的分享创意与体验的方法。图表和地图是一些早期数据可视化技术的重要例证。

2. 为什么数据可视化很重要

如上所述,人类已经使用数据可视化技术很长一段时间了,图像和图表已被证明是一种有效的方法来进行新信息的传达与教学。有研究表明,80%的人还记得他们所看到的,但只有20%的人记得他们所阅读的!它甚至可以把思想和事件传给后代。技术的发展进一步提高了数据可视化带给人们的机遇。

也许使用数据可视化的最重要的好处是它能够帮助人们更快地理解数据。你可以在

一个图表中突出显示一个大的数据量，并且人们可以快速地发现关键点。如果用书面形式，它可能需要数小时来分析所有的数据及联系。

此外，这种展示巨量数据的能力是另一个数据可视化的优点。一张图表可能会突出显示一些不同的事项，人们可以在数据上形成不同的意见。这自然能为商业开辟新的途径。人们或许能从数据中发现一些意想不到的东西。

数据的可视化展示，提高了解释信息的能力。从海量的数据和信息中寻找联系并不容易，但是图形和图表可以在几秒内提供信息。一望便知，可提供所需的信息。

以上所述，能提高在工作场所或教育机构的沟通和有效性。数据可视化被普遍认为是一种简单而有效的方法来概括数据，因此它是可以提高人们的共享信息和学习的一种方法。

6.6.2　数据可视化定义与方法

1. 数据可视化定义

数据可视化为人们提供了从阅读局部信息到纵观全局信息、从表面到本质和从内容到结构的有力工具。其演化过程是从文本到树和图，再到多媒体，以便最大限度地利用人们的多通道和分布式认知功能以及形象思维功能。

数据可视化致力于通过交互可视界面来进行分析、推理和决策。人们通过使用可视分析技术和工具，从海量、动态、不确定甚至包含相互冲突的数据中整合信息，获取对复杂情景的更深层的理解。可视分析技术允许人们对已有预测进行检验，对未知信息进行探索，提供快速、可检验和易理解的评估，以及提供更有效的交流手段。

数据可视化的开发和大部分项目开发一样，也是根据需求来根据数据维度或属性进行筛选，根据目的和用户群选用表现方式。

同一份数据可以可视化成多种看起来截然不同的形式：

- 有的可视化目标是为了观测、跟踪数据，所以就要强调实时性、变化、运算能力，可能就会生成一份不停变化、可读性强的图表。
- 有的为了分析数据，所以要强调数据的呈现度、可能会生成一份可以检索、交互式的图表。
- 有的为了发现数据之间的潜在关联，可能会生成分布式的多维的图表。
- 有的为了帮助普通用户或商业用户快速理解数据的含义或变化，会利用漂亮的颜色、动画创建生动、明了，具有吸引力的图表。
- 还有的图表可以被用于教育、宣传，被制作成海报、课件，出现在街头、广告手持、杂志和集会上。这类图表拥有强大的说服力，使用强烈的对比、置换等手段，可以创造出极具冲击力自指人心的图像。在国外许多媒体会根据新闻主题或数据，雇用设计师来创建可视化图表对新闻主题进行辅助。

2. 数据分类及可视化方法

要可视化的数据大致可分以下几类：

1）系列对象，之间相互关联

这种情况下因为要展示数据之间相互关系，所以实质上是一个网络图，不过通过一些技巧可以把简单网络图变成更好的形式。例如，转换成流图或圈形的网络图，圈形可以使得连线集中在圈内部，而且可以减少交叉。

2）层级数据

数据之间可分成几个层级关系，就是层级图。使用散点的大小或者颜色等属性来表示数据的大小。标签云也是属于此类，我们可以通过每个标签的大小颜色等等来标示数据的大小。

3）多维数据

如何将超过人类理解能力的三维以上的数据，转化为人类能视觉直观理解的可视化结果，是多维数据可视化所研究的课题。多维数据有多种传统的可视化方法，包括平行坐标、散点图矩阵和维度降维法。

4）将时间和空间可视化

通过时间的维度来查看指标值的变化情况，一般通过增加时间轴的形式，也就是常见的趋势图。

当图表存在地域信息并且需要突出表现的时候，可用地图将空间可视化，地图作为主背景呈现所有信息点。

5）让图表“动”起来

数据图形化完成后，可结合实际情况，将其变为动态化和可操控性的图表，用户在操控过程中能更好地感知数据的变化过程，提升体验。

实现动态化通常以下两种方式：交互和动画。

6）多种可视化方法结合

单一的可视化方法已不能满足需要。越来越多的可视化系统通过结合不同的科学和数据可视化方法，提供一致的多视角和连贯的交互手段，使可视化系统能够提供日益复杂的数据所需的分析能力。

3. 数据可视化常用工具

有一些用于数据可视化的工具。这些工具便于收集数据及简化数据的使用方式。一些常用工具包括：

（1）Google charts。Google 的产品在数据行业是众所周知的，Google charts 是一个方便的工具，特别是对于初次使用的用户。

（2）Datawrapper。这是一个在线工具，它可以帮助你创建交互式数据可视化。

（3）RAW。它的优点是有很多现成的模板框架让你清晰、快捷地呈现信息。该平台开源，能够自定义布局，以及使用其他的设计。

（4）Infogram。新手用户的另一个伟大工具。它允许用户创建不同的图表和信息图，而且系统易于使用。

这些都不是唯一可用的工具，你可以找到其他一些免费和付费软件。为确保你所使用的软件适合数据可视化目标，需要多多对比。

4. 数据可视化背后关键概念

看过数据可视化的人都明白设计的好坏。如果这些信息不是以正确的、恰当的方式呈现，那么数据可视化的好处就很容易消失，特定项目需要特定的方法。

无论你的信息是关于什么的，使用数据可视化时要牢记一些理念。以下是优秀数据可视化技术背后核心理念的集合。

1）了解受众

呈现数据前首要做的是思考谁将查看这些数据，为找到合适的数据可视化方法，了解受众非常关键。

尽管数据可视化通常是一种简化数据的方法，受众可能仍然存在不同的知识背景，需要为此做好准备。如果数据可视化的目标是专业受众，那么可以使用更适合的方法以及使用专业术语来解读数据。另一方面，普通受众可能需要相同的数据提供更加清晰的解释方式。

同样重要的是，要知道受众对数据的预期。他们想要的关键点是什么？你需要清楚呈现到数据中。此外，还需要明白，你的数据意图。

2）足够了解数据

除了知道你的目标受众，您还需要了解数据的内涵。如果你不完全明白你的数据，那么你将无法有效将其传达给受众。

你也无法从数据中提取所有信息，所以需要找到关键信息，并以一致的方式呈现它。还需要确定数据的正确性，错误的数据不可能可视化。

如果你正确地理解它，就可以从数据中得到独特而有趣的信息。

3）讲故事

数据可视化还应当力求传达一个故事。你不希望这些数据是一组信息仅仅呈现自己，而是有使用数据背后的信息。这可能是关于引入不同的叙述，并为观众描绘的特定图像。

使用一个故事，往往意味着受众从数据中获得更多的洞察力。它可以帮助受众了解及深入新的信息。

事实上，数据可视化技术是个讲故事的好工具。俗话说："图像可以讲述一千个故事。"这是有道理的，你应该用它来作为你的优势。通过数据集讲故事并不困难，因为你可以用颜色、字体及陈述作为你讲故事方法的一部分。

为了使数据可视化讲的故事更加精彩，理解数据这点是至关重要的。

4）保持简单

近年来，数据可视化已经发展了很快，如前所述，有很多工具和系统供你使用。接触不同的独特方法并不意味着你需要使用它们。此外，大量的数据不应该机械地认为所有的信息是必不可少的。

总之，你需要保持数据可视化方法简单明了。你不要企图让它包含太多的数据信息或使用过多不同的技术。

如果你考虑通过镜头讲故事，那么重要的是要了解你的视觉中的每个元素应该是故

事必不可少的一部分。如果数据或元素，如某些事物的图片，没有添加任何重要的故事，那么你不应该把它包含在其中。

拥有过多元素的可视化实际上会损坏成品并会偏离数据。你还需要记住数据可视化的好处是直观地呈现大量的数据。如果可视化结果看起来费劲，那么你需要回去看看是否使用了错误的数据呈现方法或包含了太多冗杂的信息。

5）正确认识平台需求

最后，一个成功的数据可视化技术也关注技术方面。现在，人们通过不同的平台查看和访问信息，重要的是你要记住这点。就像你需要知道目标受众，你也需要考虑人们阅读你的数据可视化的方式。

你需要让可视化结果方便地进行平台移植，如在移动手机、平板电脑或计算机之间移植。如果你的用户只通过手机浏览数据，那么你自然会受益于移动手机创建可视化的方法，而不是用笔记本电脑创建数据。

除了考虑该平台的界面选项外，还需要考虑可访问性问题。如果数据可视化允许有视觉障碍的人进行适当的缩放，可以大大提高用户体验。你也可以考虑不同的颜色选择供色盲者使用。可访问性有助于提高用户体验，确保你的数据可视化可用于所有受众。

5. 避免可视化数据的严重误区

以上的关键方法可以帮助你建立一个数据可视化策略，你也需要清楚一些常见的错误。

1）错误信息

上述提到数据中的错误会误导受众。你需要确保那些正在看你的数据的人，看到的信息正确。这是你的工作，以确保人们可以从你的图表和图像中使用数据，而不需要再次检查信息。

2）不完全信息

除了确保所有的信息是正确的，还需要提供完整的数据。观察者必须在其全部信息中找到相关数据，不要使用数据可视化来欺骗或呈现不完整的信息。

数据可视化可以而且应该讲述一个故事，但故事需要有完整和正确的信息，而不是一份报告中看起来合适的数字。

3）简单的数据

虽然需要确保数据是在用一个简单的方式呈现，这并不意味着简化它。首先，你需要记住受众——如果是将数据展示给专业人士就不要使用常见的简单语言。另一方面，如果受众对它没有什么认识，就不要用专业术语。

除此之外，你也不能期望受众在没有借助清晰描述的可视化形式的情况下就能清楚地了解数据之间的联系。你不能因为它似乎显而易见而省略信息——记住，受众只会看到目前的数据，而不是过去使用过的完整数据集！

4）不合适的可视化

当呈现数据时，需要仔细思考这些数据。比如字体、颜色和图像，背景也是非常重要的。例如，如果是呈现由于特定的疾病而导致死亡的信息，一个色彩鲜艳、令人愉快的图

像似乎是不合适的。

不恰当的可视化涉及所使用的技术,使它难以查看和理解数据。例如,你可以使用气泡来代表部门不同的消费水平,但如果不考虑尺寸的差异,气泡就会误判和不准确。

5）遗忘注释

过度简化也可能导致缺失注释。在呈现数据时,很容易假设受众知道图像的每一个方面是什么。简单地添加注释可以提高用户体验,并确保受众知道数据中的所有数据关键点。

作为一个例子,你可能有一个图表显示企业在过去十年销售自行车量。如果数据中有一个大的下降或是上升,一个注释解释了这个突然变化背后的原因,将确保观众得到这个额外的信息。

6. 信息可视化案例

信息可视化囊括了数据可视化、信息图形、知识可视化、科学可视化以及视觉设计方面的所有发展与进步。下面是信息可视化的案例分享。

关系网——基于60 000封电子邮件存档数据,用不同颜色深度的线条呈现了地址簿中用户和个体之间的关系,比如回复、发送、抄送。

关系网的信息可视化如图6.14所示。

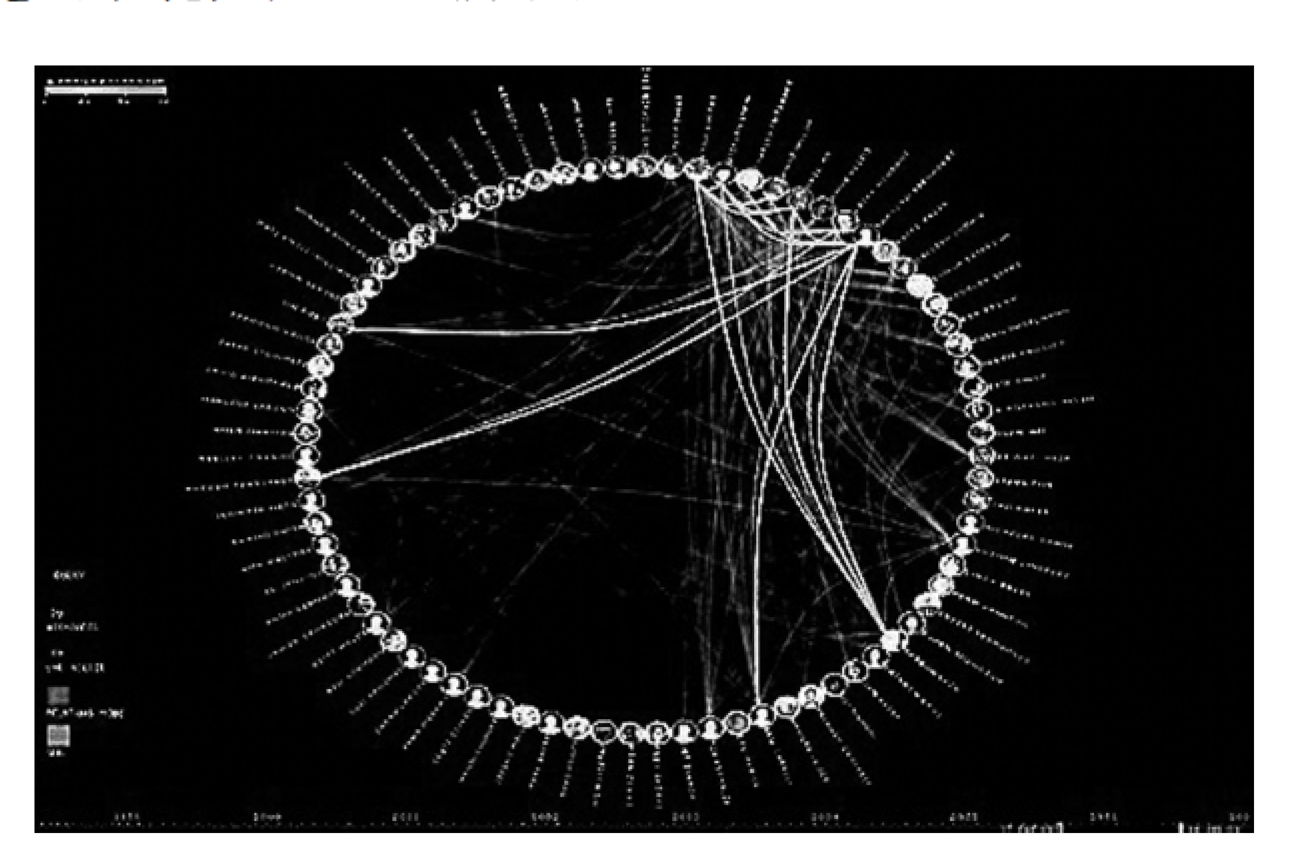


图6.14 关系网

根据ESM国际电子商情针对大数据应用现状和趋势的调查显示：被调查者最关注的大数据技术中,排在前五位的分别是大数据分析(12.91%)、云数据库(11.82%)、Hadoop(11.73%)、内存数据库(11.64%)以及数据安全(9.21%)。

既然大数据分析是最被关注的技术趋势,那么大数据分析中的哪项功能是最重要的呢?研究发现,排在前三位的功能分别是实时分析(21.32%)、丰富的挖掘模型(17.97%)和可视化界面(15.91%)。企业对实时分析的需求激增,成就了很多以实时分析为创新技术的大数据厂商。

从调查结果可以看出：企业在未来一两年中有迫切部署大数据的需求,并且已经从一开始的基础设施建设,逐渐发展为对大数据分析和整体大数据解决方案的需求。我们一起看看以下哪五大类数据产品有大数据应用的踪影。

6.6.3 数据可视化分析

数据可视化分析一般可以分为以下几种类型。

1. 原始数据分析

有时客户并不完全了解自己的数据，人员更替、平台迁移、数据遗失、没有专门的负责人去进行数据的管理和维护，都会造成数据的资源浪费。虽然随着时间过去，越早的数据价值越小，但是有人说过，不能坦然面对过去的人，也无法面对将来。所以，先从整理过去开始吧。

2. 营销数据分析

营销数据的重要性就不用赘述，既要多纬度多，又要分析深刻结论明了。最好是又美观又能方便导出，还可以通过邮箱分享或者嵌入网页。

3. 业务场景数据分析

能把已有业务场景数据可视化是比较个性化的需求了，但是一旦实现出来，在某种程度上说还是能增加工作效率，如图 6.15 所示。



图 6.15 业务场景数据分析

一些例子表明，可视化是有助于监控风险。

银行客户订制了一套基于转账的可视化系统，若有人打款，就会从打款地发出一条光束到达收款地。就在管理层观察了一段时间后惊人的发现，在每天的同一时间段，有 100 多条光束会同时汇集落到同一地点，也就是说，100 多个账户在打款进同一账户中。最后经过查证，是不法行为。这就是通过数据可视化直观监测反洗钱的典型案例。

4. 地理位置数据分析

一般的 LBS 场景是，将业务数据放置于地图中，用户可以获取可视化的数据分析，并能自行上传位置数据。但是现在也有结合物联网需求的可视化地理位置分析，是不是更有实感？看见我的快递在努力地朝我的方向移动，突然有点感动……

5. 用户画像

当某人真的被准确地定位成“女屌丝”的那一刻，就会发现，她或许不太喜欢这个功能。所以并不面向用户本身的话，可能还不错。让商家去具象地了解用户的信息，做出判

断和营销。

用户画像如图 6.16 所示。



图 6.16　用户画像

6.6.4　个性化精准推荐

下一波的数字化淘金浪潮将会是如何利用数据来解决实际问题，而不仅仅是使用数据的行为。“未来已经来临，只是尚未流行”—— 著名研究机构 Gartner。

在技术不到位、数据储备不足的情况下，个性化服务可能出力不讨好。理论上个性化服务可以消除通知噪声来提高现有用户满意度，同时可以发展新用户，利用长尾效应增加收益。

1. 订阅推荐

订阅选项真的非常丰富。或关联社交账户，或通过搜索关注话题，或根据以往阅读文章推论，或根据关注对象……订阅推荐如图 6.17 所示。

2. 商品推荐

根据你浏览过的推荐，根据你购买过的推荐，根据和你一样购买过的人推荐，虽说老套，但成功率也高。商品推荐如图 6.18 所示。

3. 社交图谱 & 兴趣图谱

社交图谱 & 兴趣图谱把所有和你有关的都连在一起。在很多企业中，社交图谱分析已经在反欺诈、影响力分析、舆情监测、市场细分、参与优化、体验优化，以及其他需要快速确定复杂行为模式的领域成功应用。社交图谱与兴趣图谱如图 6.19 所示。

当我知道我看到的这个东西是完完全全为我打造的时候，我更想知道，别人在看些啥……我上网就是为了融入这个世界啊。

6.6.5　预测和预警

预测和预警无论是在商业或者是生活问题解决上都是有实际意义的，在初期，人们对其可到达的精准程度还是有一定担忧。但是播了几十年的天气预报也不是很准啊……

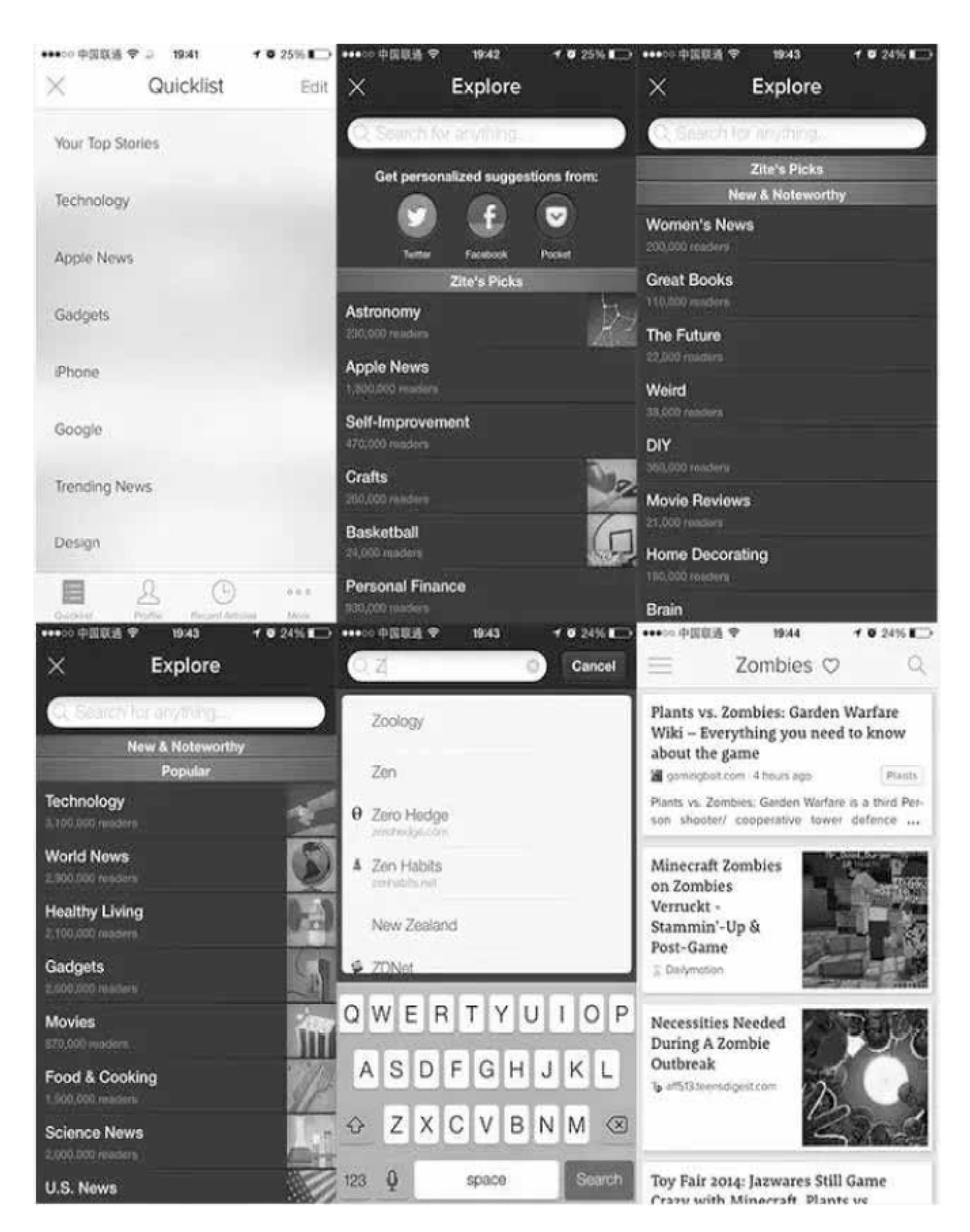


图 6.17　订阅推荐

☆ 与您浏览过的商品相关的推荐



图 6.18　商品推荐

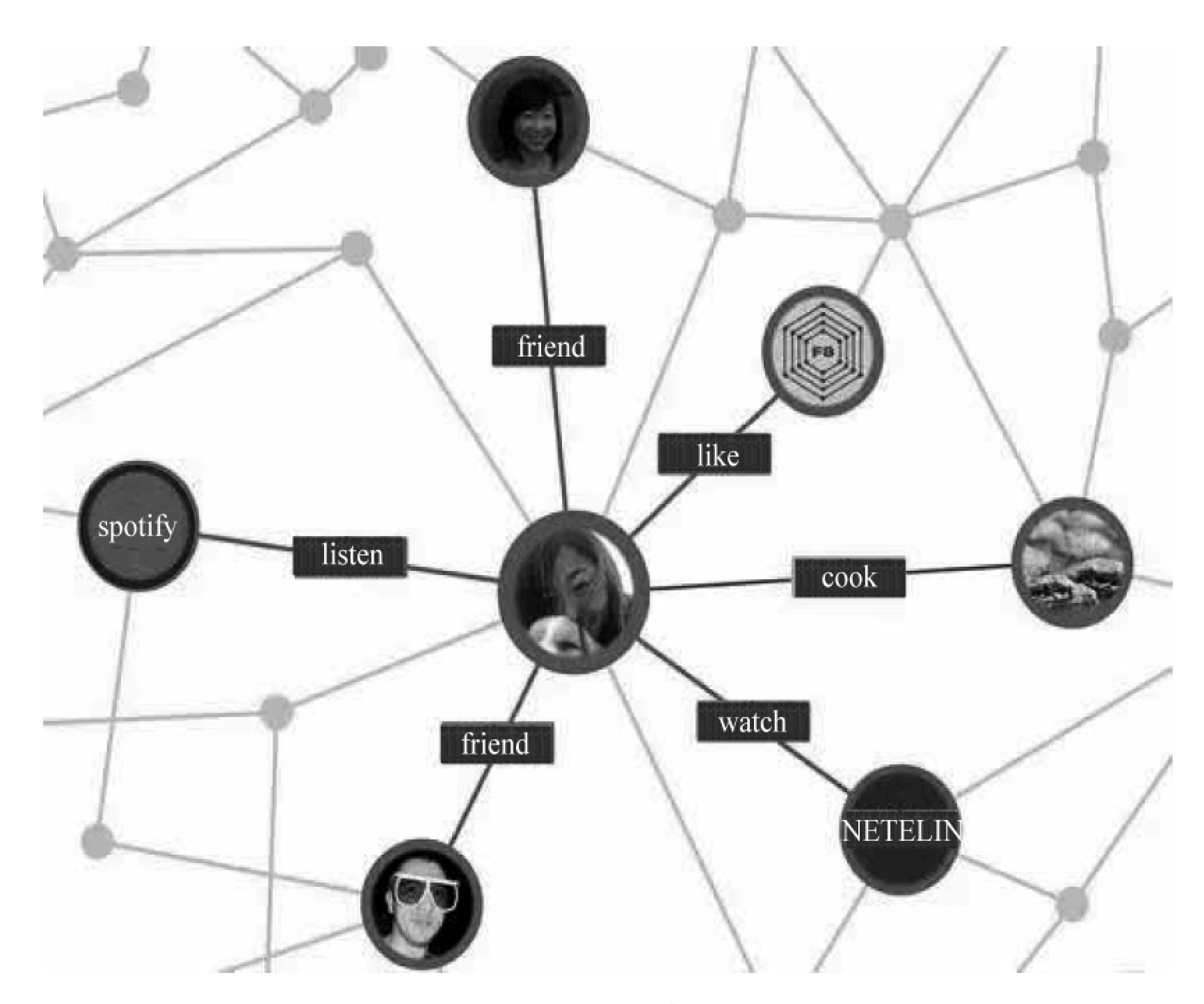


图 6.20　社交图谱 & 兴趣图谱

1. 交通状况预测

监控提供的数据可以帮助追踪道路交通情况，可以进行线路推荐和目的地到达时间的预测。通过算法，如果街道上涌现出大批人群，车辆可以及时进行交通道路调整。

2. 医疗类预测

利用数据库中病情发展记录做出预测。这种预测将基于对患者日常行为的观测，力求在病情出现恶化之间就介入治疗。甚至有机构调查一些拥有长寿者的家谱和基因里蕴含的生命信息。最后即使不能通过研究找到延长寿命的方法，但至少能通过疾病预防，提高老年群体的生活质量。医疗类预测如图 6.20 所示。

3. 消费信誉预测

通过数据挖掘分析和机器学习技术，对申请者提交的信息进行识别，并结合个人社交行为及海量互联网信息，对个人信用进行在线评分。基于强大的数据点基础，很快让用户得到信用额度，额度可以用在各类金融和非金融服务领域。

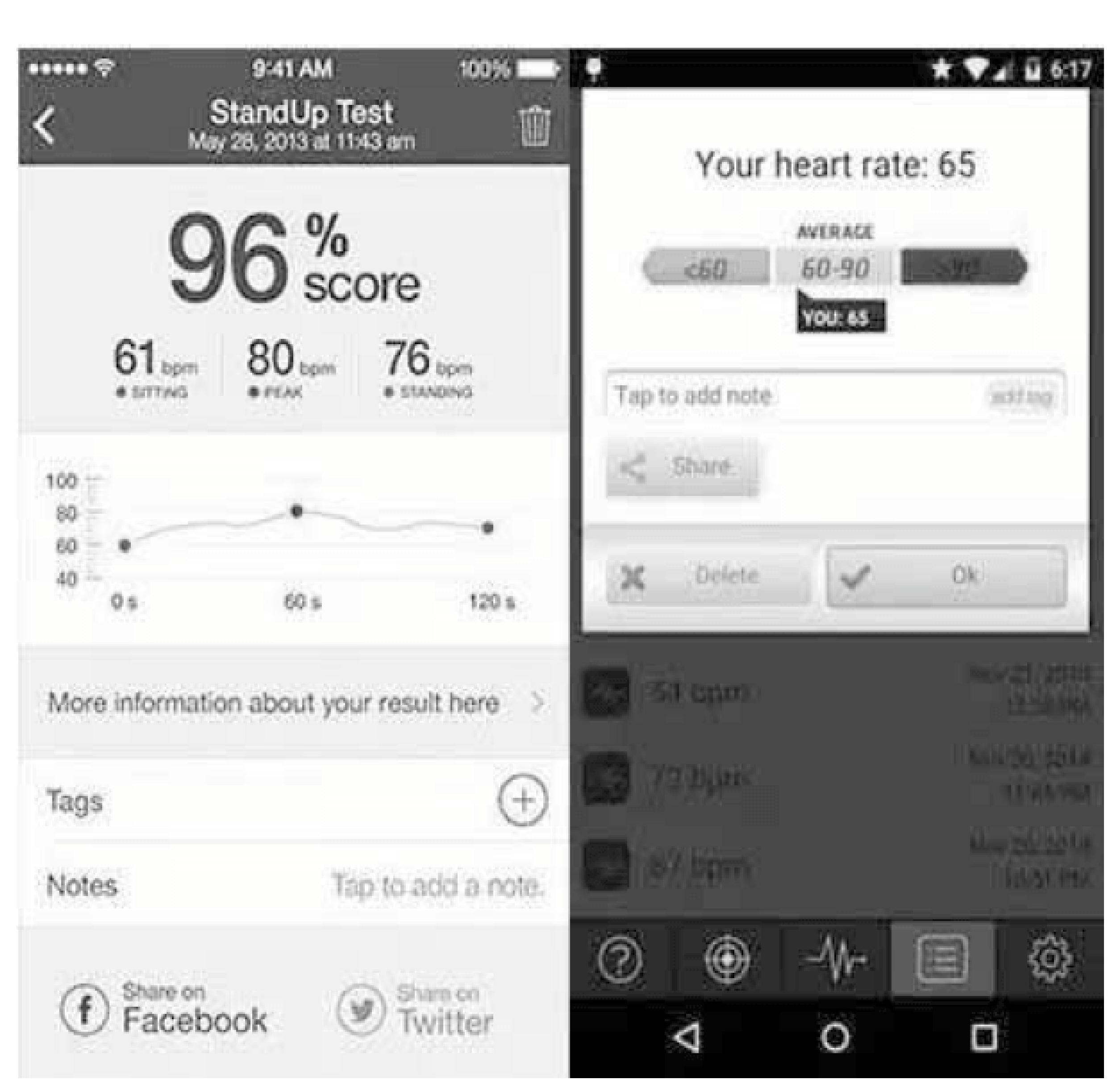
消费信誉预测如图 6.21 所示。

6.6.6　决策分析

大到总金额无法计算的商业决策，小到站在包子铺门口的纠结、出门走哪条路、参加朋友婚礼穿什么衣服，若是真有完美的决策分析，无疑是选择恐惧症患者的福音。

1. 销售决策

比如一个购物网站，当消费者登录这个网站时，会把这名消费者在网站上的行为和以



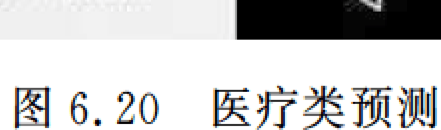

图 6.20　医疗类预测



图 6.21　消费信誉预测

前其他登录过该网站的消费者行为做对比，做出分析和预测，然后给出一份实时的建议：例如，现在平台是应该向消费者抛出一个聊天信息、一个产品打折的报价、一个视频对话、还是一个电话会比较好？——或者是什么都不做最好。

2. 旅行决策

通过抓取海量数据，分析提取关键字、建立评分体系，让用户不用看长篇攻略就能掌握核心信息，快速做出旅行决策。

对于大数据的定义，著名研究机构 Gartner 给出了这样的定义：“大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。”去掉这句话里所有的定语，得到的是：大数据是信息资产。所以，我们知道了，不管有没有大到哪一种体量级别，至少让数据信息成为一种资产也算是有大数据精神了。

6.7 知识图谱

知识图谱（Knowledge Graph）是当前的研究热点。自从 2012 年 Google 推出自己第一版知识图谱以来，它在学术界和工业界掀起了一股热潮。各大互联网企业在之后的短短一年内纷纷推出了自己的知识图谱产品以作为回应。比如在国内，互联网巨头百度和搜狗分别推出“知心”和“知立方”来改进其搜索质量。

6.7.1　知识图谱的概念

知识图谱本质上是语义网络，是一种基于图的数据结构，由结点(Point)和边(Edge)组成。在知识图谱里，每个结点表示现实世界中存在的"实体"，每条边为实体与实体之间的"关系"。知识图谱是关系的最有效的表示方式。语义网络如图 6.22 所示。

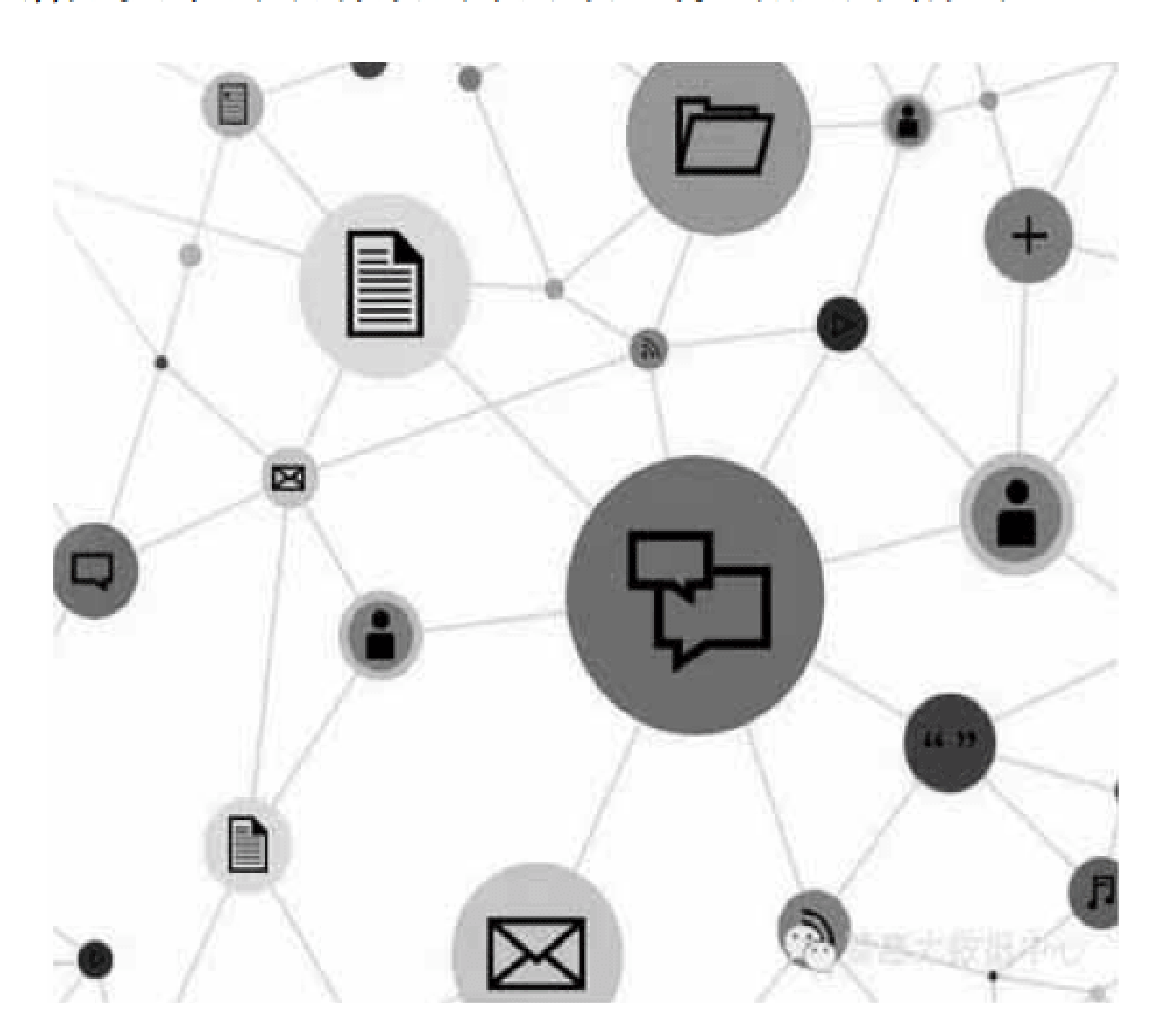


图 6.22　语义网络

通俗地讲，知识图谱就是把所有不同种类的信息(Heterogeneous Information)连接在一起而得到的一个关系网络。知识图谱提供了从"关系"的角度去分析问题的能力。

知识图谱这个概念最早由 Google 提出，主要是用来优化现有的搜索引擎。不同于基于关键词搜索的传统搜索引擎，知识图谱可用来更好地查询复杂的关联信息，从语义层面理解用户意图，改进搜索质量。比如在 Google 的搜索框里输入 Bill Gates 的时候，搜索结果页面的右侧还会出现 Bill Gates 相关的信息比如出生年月、家庭情况等等。

另外，对于稍微复杂的搜索语句比如"Who is the wife of Bill Gates"，Google 能准确返回他的妻子 Melinda Gates 的信息。这就说明搜索引擎通过知识图谱真正理解了用户的意图。

上面提到的知识图谱都属于比较宽泛的范畴，在通用领域里解决搜索引擎优化和问答系统(Question-Answering)等方面的问题。接下来介绍特定领域里的(Domain-Specific)知识图谱表示方式和应用，这也是工业界比较关心的话题。

6.7.2　知识图谱的表示

假设我们用知识图谱来描述一个事实(Fact)——"张三是李四的父亲"。这里的实体是张三和李四，关系是"父亲"(is_father_of)。当然，张三和李四也可能会跟其他人存在着某种类型的关系(暂时不考虑)。当我们把电话号码也作为结点加入到知识图谱以后(电话号码也是实体)，人和电话之间也可以定义一种关系叫 has_phone，就是说某个电话号码是属于某个人。图 6.23 就展示了这两种不同的关系。

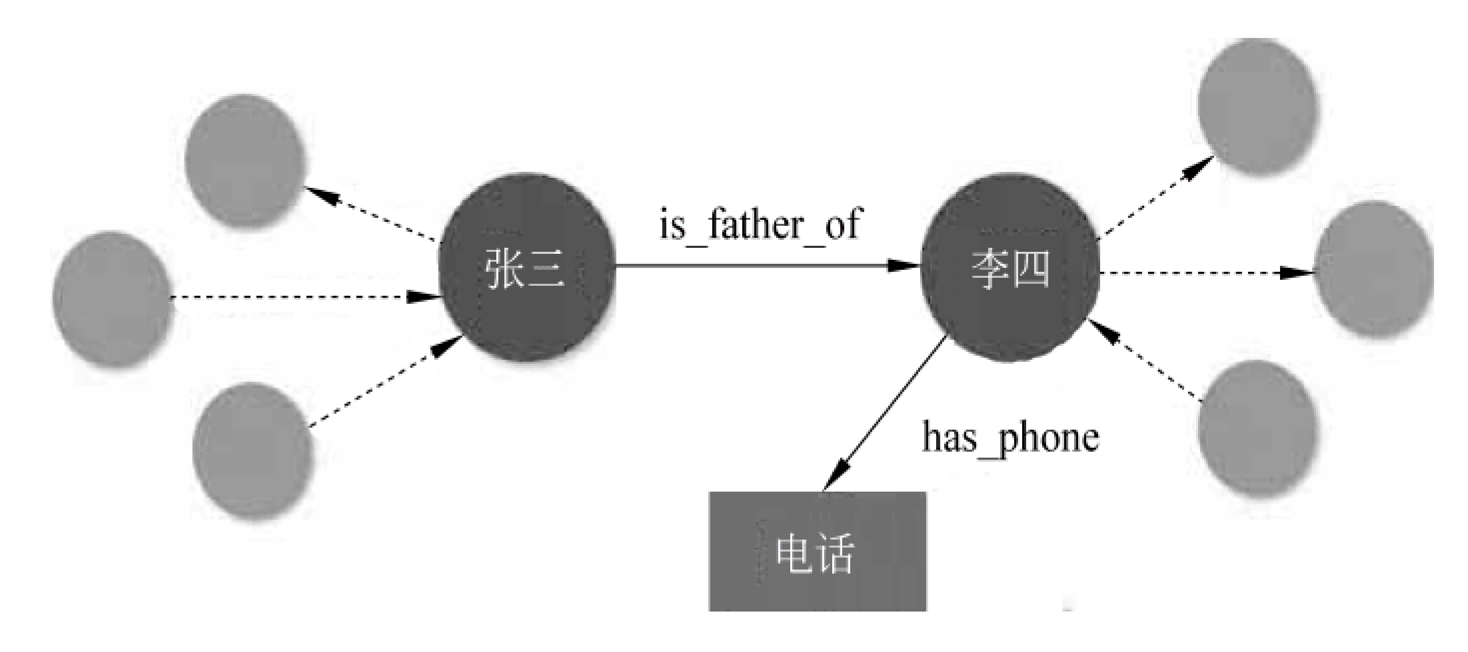


图 6.23 事实(Fact)——"张三是李四的父亲"

另外,我们可以把时间作为属性(Property)添加到 has_phone 关系里来表示开通电话号码的时间。这种属性不仅可以加到关系里,还可以加到实体当中,当我们把所有这些信息作为关系或者实体的属性添加后,所得到的图谱称为属性图(Property Graph)。属性图和传统的 RDF 格式都可以作为知识图谱的表示和存储方式,但二者还是有区别的,这将在后面做简单说明。

6.7.3 知识图谱的存储

知识图谱是基于图的数据结构,它的存储方式主要有两种形式:RDF 存储格式和图数据库(Graph Database)。

如表 6.5 所示的是目前比较流行的基于图存储的数据库排名。从这个排名中可以看出,Neo4j 在整个图存储领域里占据着 NO.1 的地位,而且在 RDF 领域里 Jena 还是目前为止最为流行的存储框架。

表 6.5 流行的基于图存储的数据库排名

Ranking	DBMS	Ranking	DBMS
21	Neo4j(图)	61	Virtuoso(RDF,关系等)
32	MarkLogic(XML)	80	Jena(RDF)
42	Titan(图)	88	Sesame(RDF)
46	OrientDB(图,文档)	90	ArangoDB(图)

当然,如果需要设计的知识图谱非常简单,而且查询也不会涉及 1 度以上的关联查询,那么也可以选择用关系型数据存储格式来保存知识图谱。但对那些稍微复杂的关系网络(现实生活中的实体和关系普遍都比较复杂),知识图谱的优点还是非常明显的。首先,在关联查询的效率上会比传统的存储方式有显著的提高。当涉及 2、3 度的关联查询时基于知识图谱的查询效率会高出几千倍甚至几百万倍。其次,基于图的存储在设计上会非常灵活,一般只需要局部的改动即可。比如有一个新的数据源,只需要在已有的图谱上插入就可以。与此相反,关系型存储方式灵活性方面比较差,它所有的 Schema 都是提前定义好的,如果后续要改变,那么代价是非常高的。最后,把实体和关系存储在图数据结构是一种符合整个故事逻辑的最好的方式。

6.7.4　知识图谱的应用

以下主要讨论知识图谱在互联网金融行业中的应用，当然，很多应用场景和想法都可以延伸到其他行业。这里提到的应用场景只是冰山一角，在很多其他的应用上，知识图谱仍然可以发挥它潜在的价值。

1. 反欺诈

反欺诈是风控中非常重要的一道环节。基于大数据的反欺诈的难点在于如何把不同来源的数据(结构化、非结构)整合在一起，并构建反欺诈引擎，从而有效地识别出欺诈案件(比如身份造假、团体欺诈、代办包装等)。不少欺诈案件会涉及复杂的关系网络，这也给欺诈审核带来了新的挑战。作为关系的直接表示方式，知识图谱可以很好地解决这两个问题。首先，知识图谱提供非常便捷的方式来添加新的数据源，这一点在前面提到过；其次，知识图谱本身就是用来表示关系的，这种直观的表示方法可以帮助我们更有效地分析复杂关系中存在的特定的潜在风险。

反欺诈的核心是人，首先需要把与借款人相关的所有的数据源打通，并构建包含多数据源的知识图谱，从而整合成为一台机器可以理解的结构化的知识。在这里，我们不仅可以整合借款人的基本信息(比如申请时填写的信息)，还可以把借款人的消费记录、行为记录、网上的浏览记录等整合到整个知识图谱里，从而进行分析和预测。这里的一个难点是很多的数据都是从网络上获取的非结构化数据，需要利用机器学习、自然语言处理技术把这些数据变成结构化的数据，如图 6.24 所示。

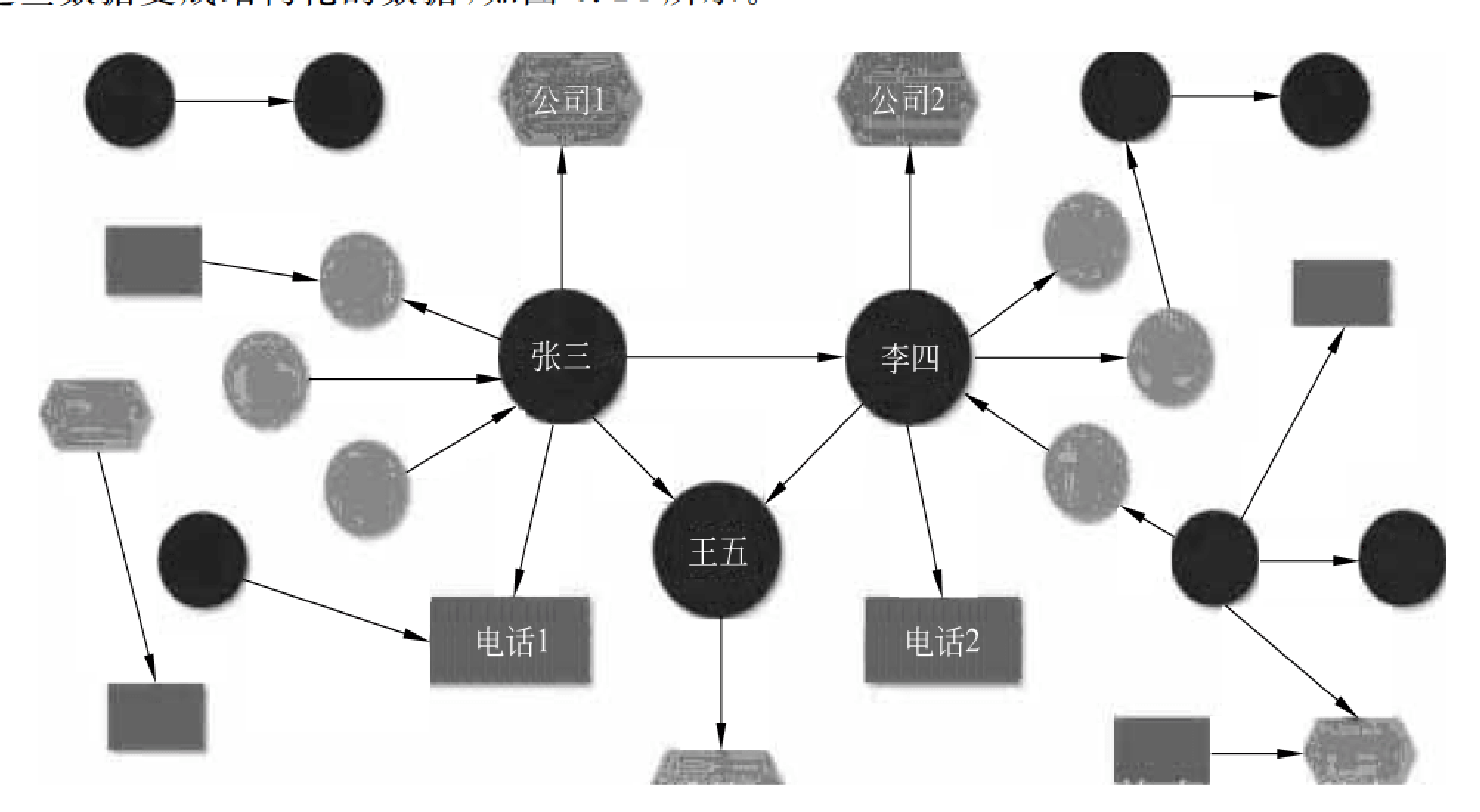


图 6.24　反欺诈

2. 不一致性验证

不一致性验证可以用来判断一个借款人的欺诈风险，这个跟交叉验证类似。比如借款人张三和借款人李四填写的是同一个公司电话，但张三填写的公司和李四填写的公司完全不一样，这就成了一个风险点，需要审核人员格外注意，如图 6.25 所示。

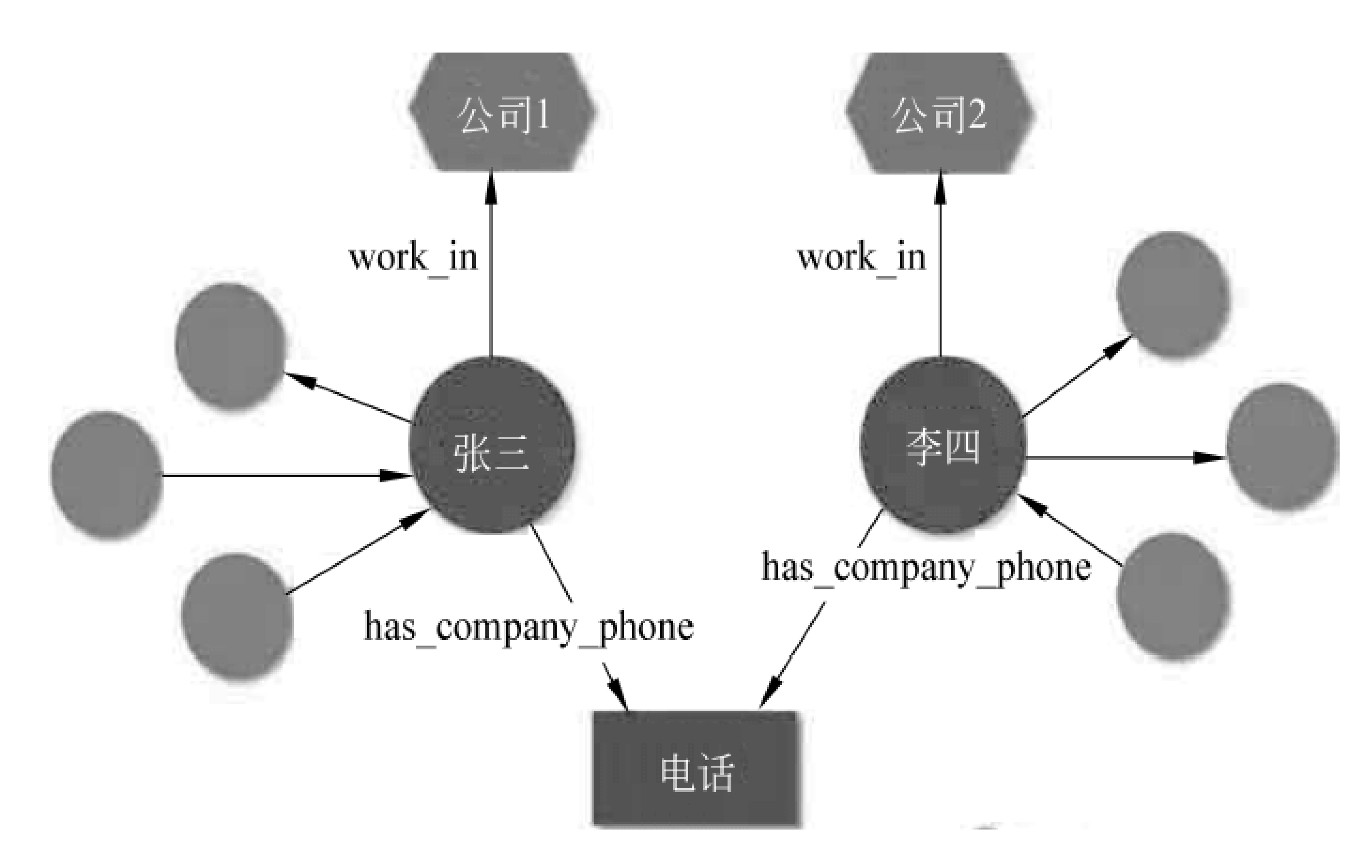


图 6.25 不一致性验证

再比如,借款人说跟张三是朋友关系,跟李四是父子关系。当我们试图把借款人的信息添加到知识图谱里的时候,"一致性验证"引擎会触发。引擎首先会去读取张三和李四的关系,从而去验证这个"三角关系"是否正确。很显然,朋友的朋友不是父子关系,所以存在着明显的不一致性,如图 6.26 所示。

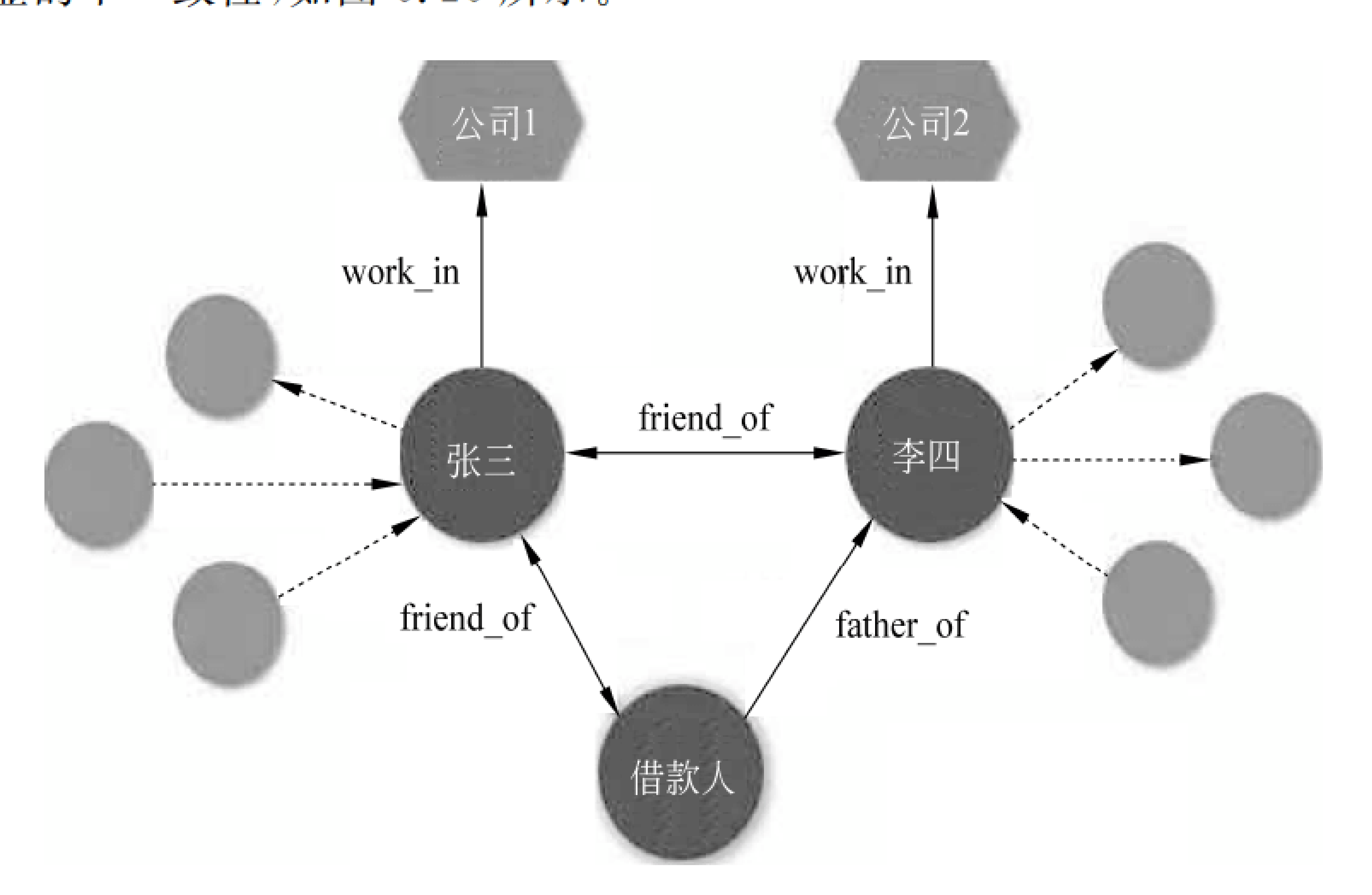


图 6.26 存在着明显的不一致性

不一致性验证涉及知识的推理。通俗地讲,知识的推理可以理解成"链接预测",也就是从已有的关系图谱里推导出新的关系或链接。比如在上面的例子,假设张三和李四是朋友关系,而且张三和借款人也是朋友关系,那我们可以推理出借款人和李四也是朋友关系。

3. 组团欺诈

相比虚假身份的识别,组团欺诈的挖掘难度更大。这种组织在非常复杂的关系网络里隐藏着,不容易被发现。当我们只有把其中隐含的关系网络梳理清楚,才有可能去分析并发现其中潜在的风险。知识图谱,作为天然的关系网络的分析工具,可以帮助我们更容易地去识别这种潜在的风险。举一个简单的例子,有些组团欺诈的成员会用虚假的身份

去申请贷款,但部分信息是共享的。图 6.27 大概说明了这种情形。从图中可以看出张三、李四和王五之间没有直接的关系,但通过关系网络我们很容易看出这三者之间都共享着某一部分信息,这就让我们马上联想到欺诈风险。虽然组团欺诈的形式众多,但有一点值得肯定的是知识图谱比其他任何的工具都更能提供更加便捷的分析手段。

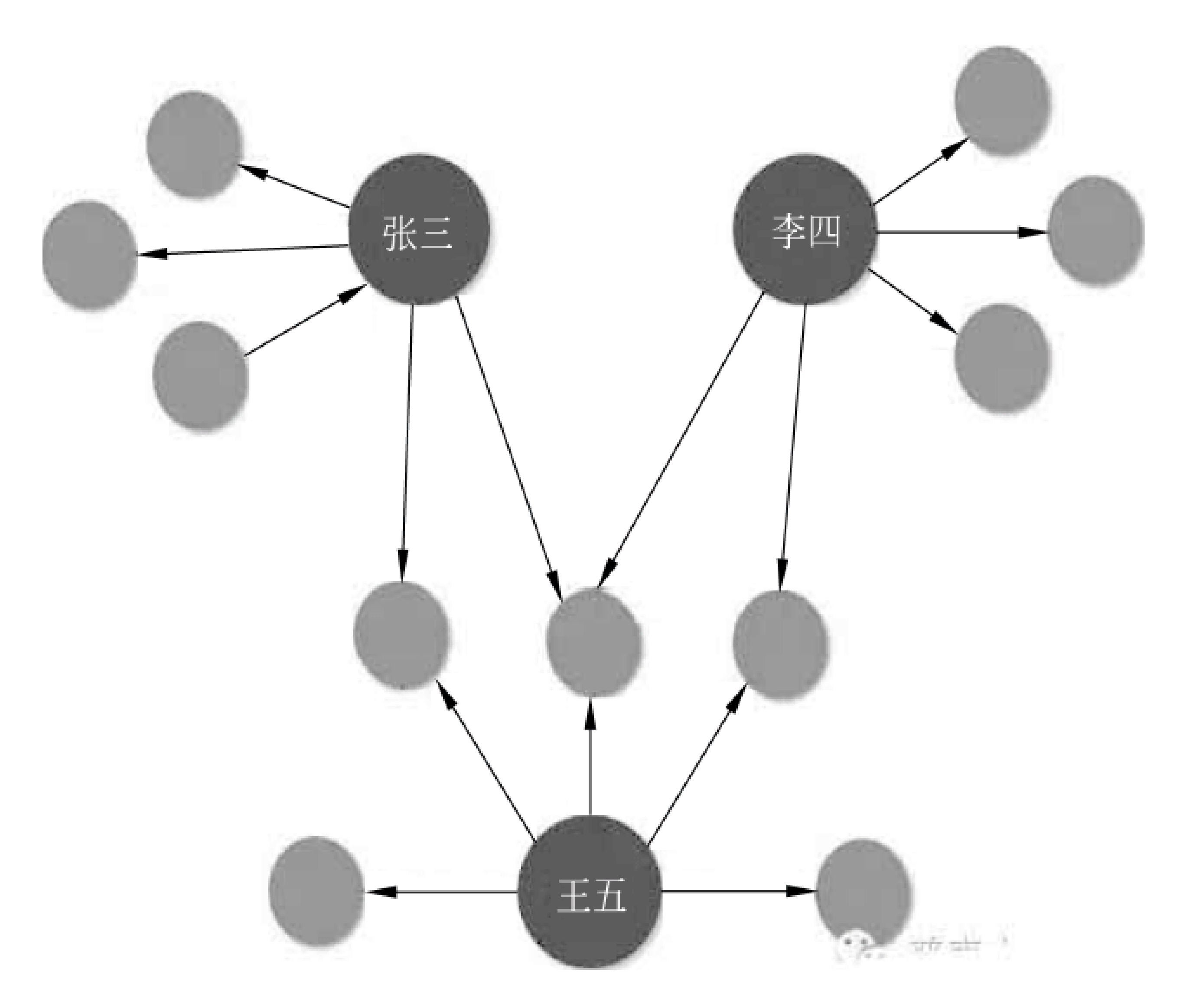


图 6.27　组团欺诈

4. 异常分析(Anomaly Detection)

异常分析是数据挖掘研究领域里比较重要的课题。我们可以把它简单理解成从给定的数据中找出"异常"点。在应用中,这些"异常"点可能会关联到欺诈。既然知识图谱可以看做是一个图(Graph),知识图谱的异常分析也大都是基于图的结构。由于知识图谱里的实体类型、关系类型不同,异常分析也需要把这些额外的信息考虑进去。大多数基于图的异常分析的计算量比较大,可以选择做离线计算。在应用框架中,可以把异常分析分为两大类:静态分析和动态分析。

1) 静态分析

所谓的静态分析,是指给定一个图形结构和某个时间点,从中去发现一些异常点(比如有异常的子图)。在图 6.28 中可以很清楚地看到其中五个点的相互紧密度非常强,可能是一个欺诈组织。所以针对这些异常的结构,我们可以做出进一步的分析。

2) 动态分析

所谓的动态分析,是指分析其结构随时间变化的趋势。我们的假设是,在短时间内知识图谱结构的变化不会太大,如果它的变化很大,就说明可能存在异常,需要进一步关注。分析结构随时间的变化会涉及时序分析技术和图相似性计算技术。有兴趣的读者可以去参考这方面的资料,如图 6.29 所示。

3) 失联客户管理

除了贷前的风险控制,知识图谱也可以在贷后发挥其强大的作用。比如在贷后失联客户管理的问题上,知识图谱可以帮助我们挖掘出更多潜在的新的联系人,从而提高催收

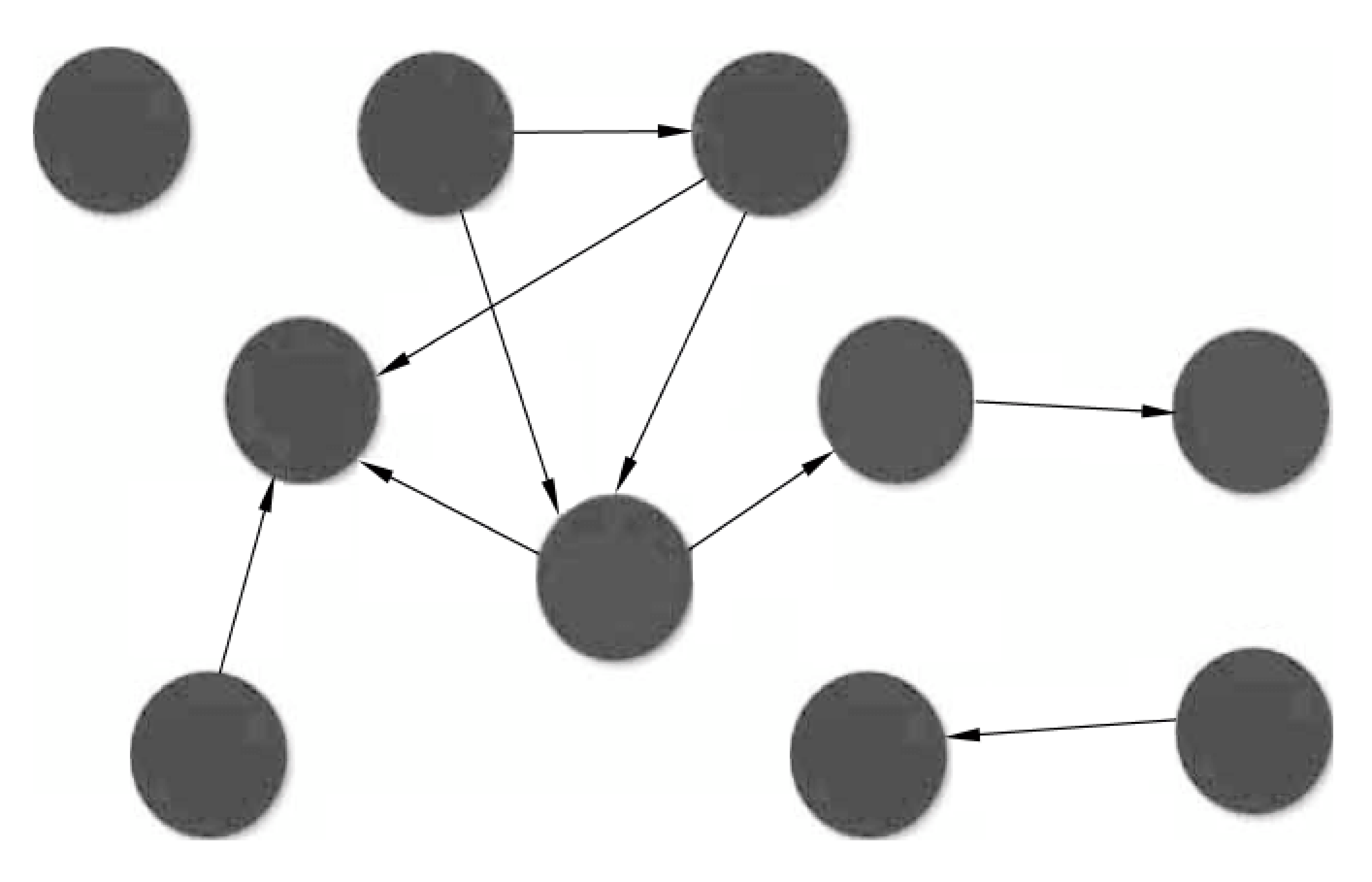


图 6.28 静态分析

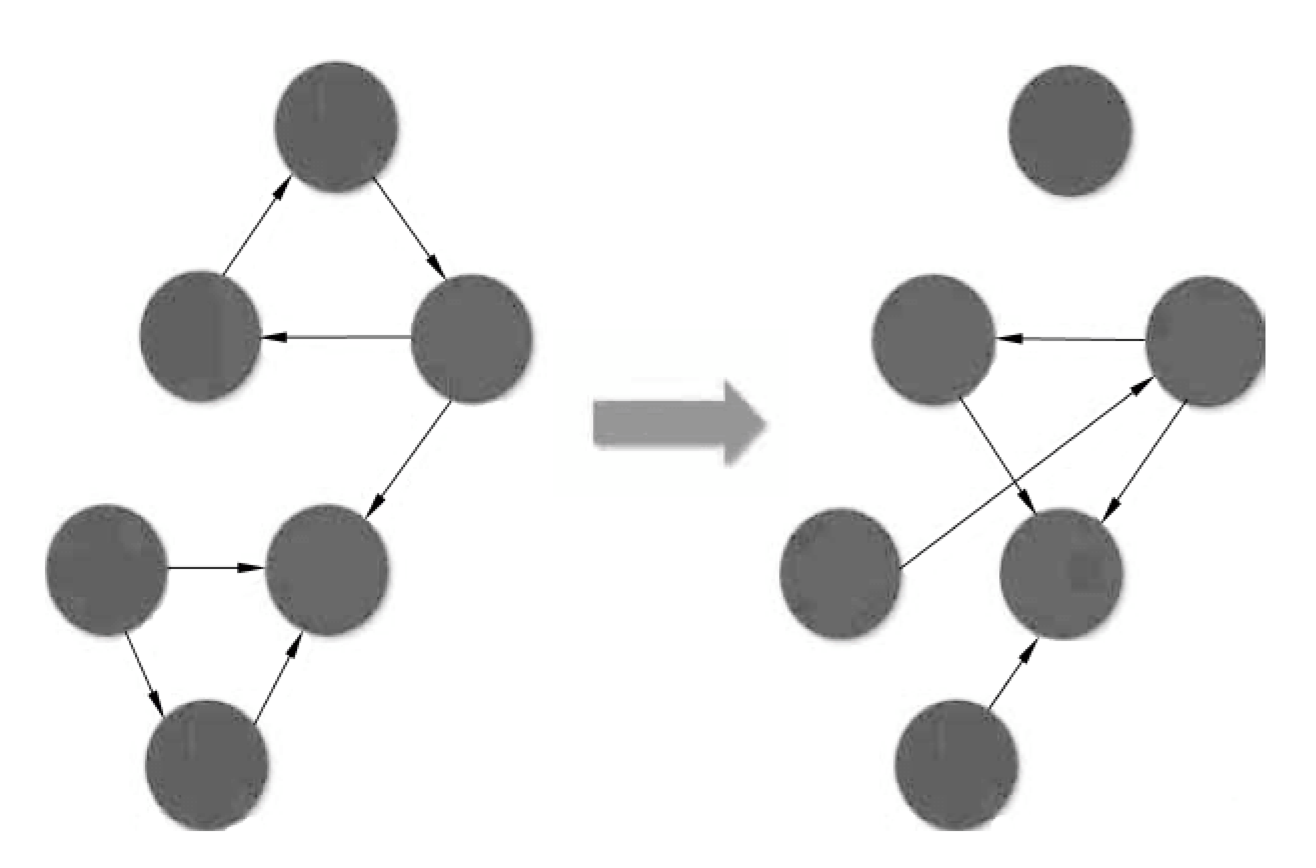


图 6.29 动态分析

的成功率。

现实中，不少借款人在借款成功后出现不还款现象，而且玩“捉迷藏”，联系不上本人。即便试图去联系借款人曾经提供过的其他联系人，但还是没有办法联系到本人。这就进入了所谓的“失联”状态，使得催收人员也无从下手。那接下来的问题是，在失联的情况下，我们有没有办法去挖掘跟借款人有关系的新的联系人？而且这部分人群并没有以关联联系人的身份出现在我们的知识图谱里。如果能够挖掘出更多潜在的新的联系人，就会大大地提高催收成功率。举个例子，在如图 6.30 所示的关系图中，借款人跟李四有直接的关系，但我们却联系不上李四。那有没有可能通过 2 度关系的分析，预测并判断哪些李四的联系人可能会认识借款人。这就涉及图谱结构的分析。

4）智能搜索及可视化展示

基于知识图谱，我们也可以提供智能搜索和数据可视化的服务。智能搜索的功能类似于知识图谱在 Google、百度上的应用。也就是说，对于每一个搜索的关键词，我们可以通过知识图谱来返回更丰富、更全面的信息。比如搜索一个人的身份证号，我们的智能搜索引擎可以返回与这个人相关的所有历史借款记录、联系人信息、行为特征和每一个实体的标签（比如黑名单、同业等）。另外，可视化的好处不言而喻，通过可视化把复杂的信息

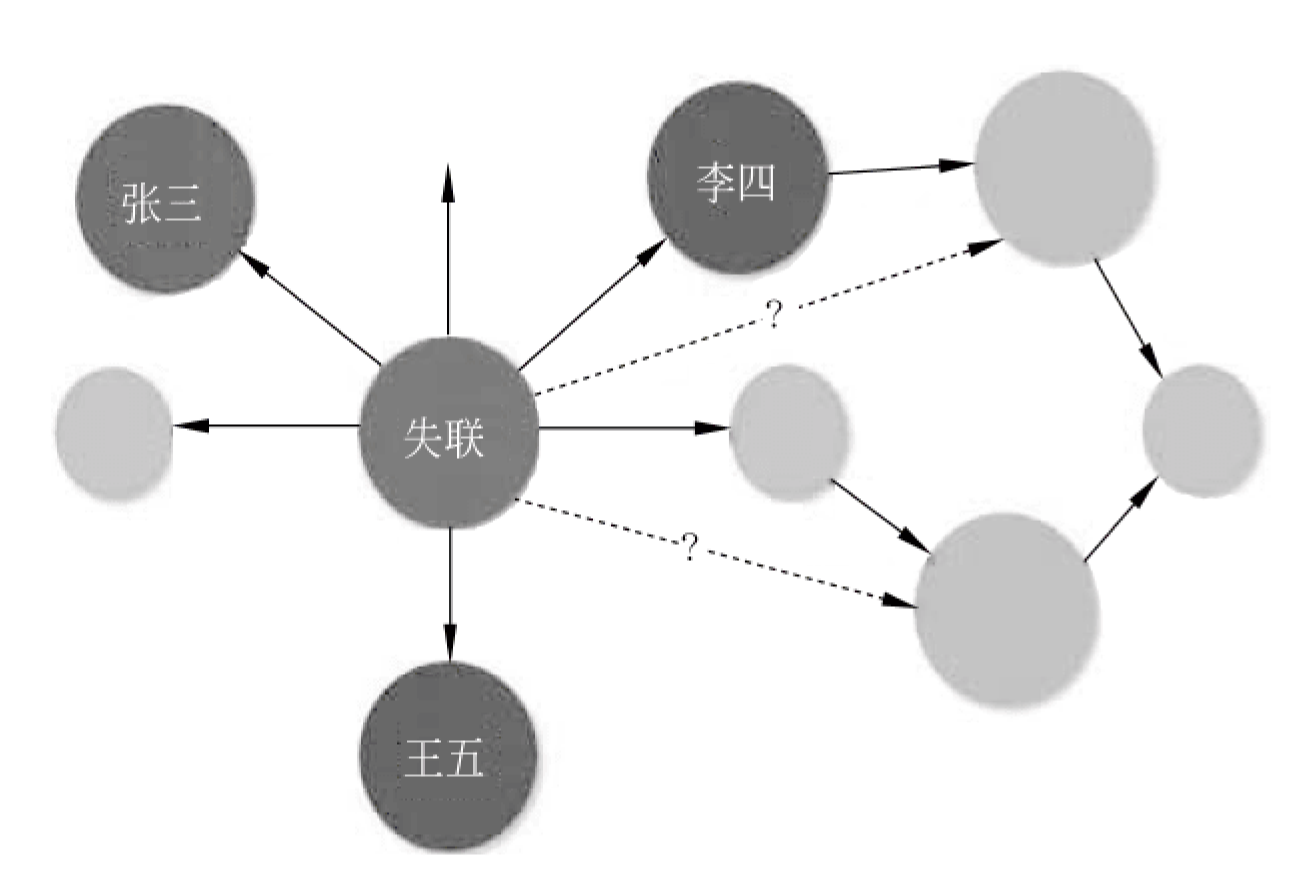


图 6.30　失联客户管理

以非常直观的方式呈现出来，使得我们对隐藏信息的来龙去脉一目了然。

5）精准营销

一个聪明的企业可以比它的竞争对手以更为有效的方式去挖掘其潜在的客户。在互联网时代，营销手段多种多样，但不管有多少种方式，都离不开一个核心——分析用户和理解用户。知识图谱可以结合多种数据源去分析实体之间的关系，从而对用户的行为有更好的理解。比如一个公司的市场经理用知识图谱来分析用户之间的关系，去发现一个组织的共同喜好，从而可以有针对性地对某一类人群制定营销策略。只有能更好地、更深入地（Deep understanding）理解用户的需求，才能更好地去做营销。

5. 挑战

知识图谱在工业界还没有形成大规模的应用。即便有部分企业试图往这个方向发展，但很多仍处于调研阶段。主要的原因是很多企业对知识图谱并不了解，或者理解不深。但有一点可以肯定的是，知识图谱在未来几年内必将成为工业界的热门工具，这也是从目前的趋势中很容易预测到的。当然，知识图谱毕竟是一个比较新的工具，所以在实际应用中一定会涉及或多或少的挑战。

1）数据的噪声

首先，数据中存在着很多的噪声。即便是已经存在库里的数据，我们也不能保证它有100%的准确性。在这里主要从两个方面说起。

第一，目前积累的数据本身有错误，所以这部分错误数据需要纠正。最简单的纠正办法就是做离线的不一致性验证。

第二，数据的冗余。比如借款人张三填写公司名字为“普惠”，借款人李四填写的名字为“普惠金融”，借款人王五则填写成“普惠金融信息服务有限公司”。虽然这三个人都隶属于一家公司，但由于他们填写的名字不同，计算机则会认为他们三个是来自不同的公司。那接下来的问题是，怎么从海量的数据中找出这些存在歧义的名字并将它们合并成一个名字？这就涉及自然语言处理中的“消歧分析”技术，如图 6.31 所示。

2）非结构化数据处理能力

在大数据时代，很多数据都是未经处理过的非结构化数据，比如文本、图片、音频、视

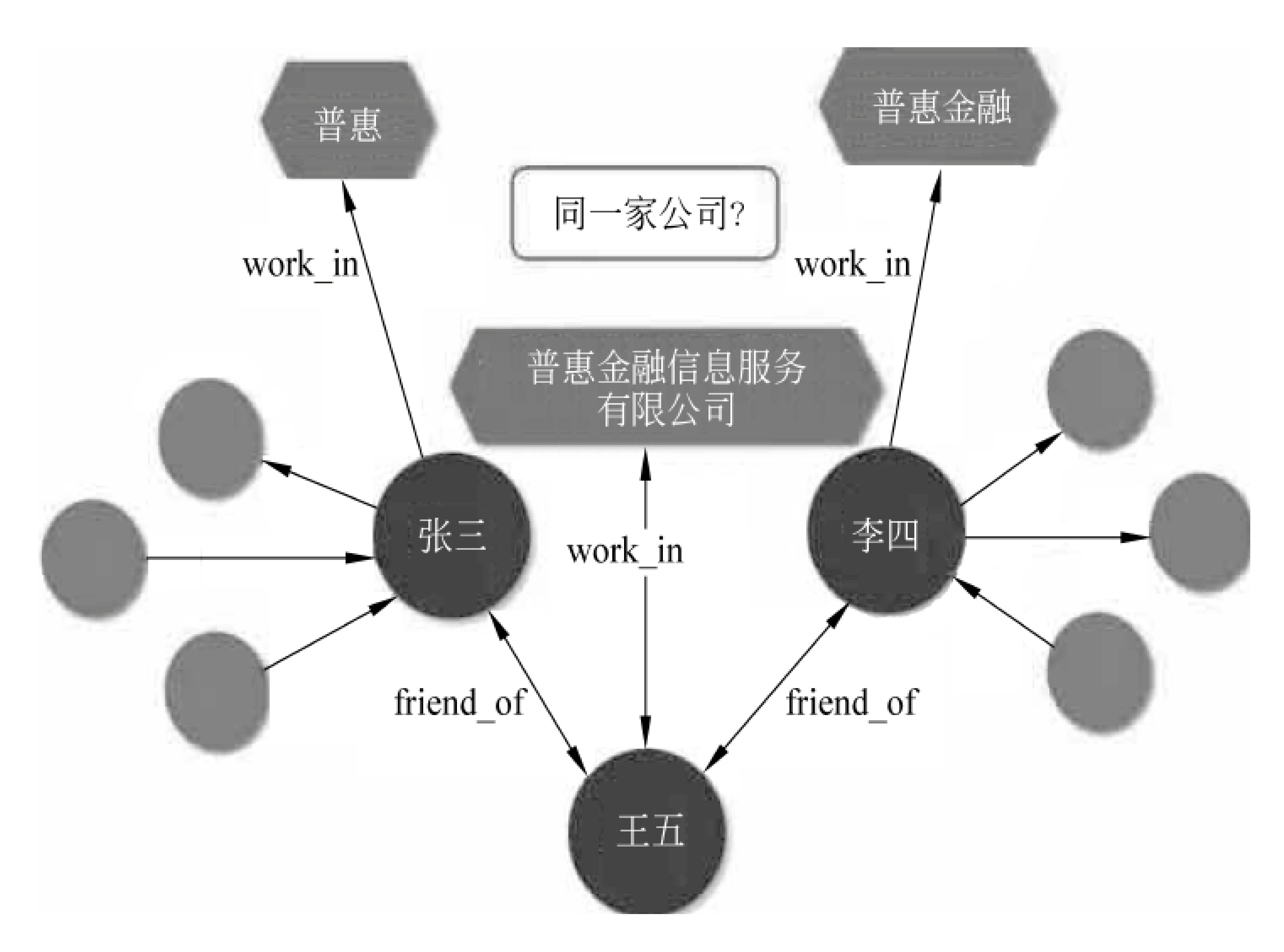


图 6.31 数据的噪声

频等。特别在互联网金融行业里,我们往往会面对大量的文本数据。怎么从这些非结构化数据里提取出有价值的信息是一件非常有挑战性的任务,这对我们所掌握的机器学习、数据挖掘、自然语言处理能力提出了更高的要求,如图 6.32 所示。

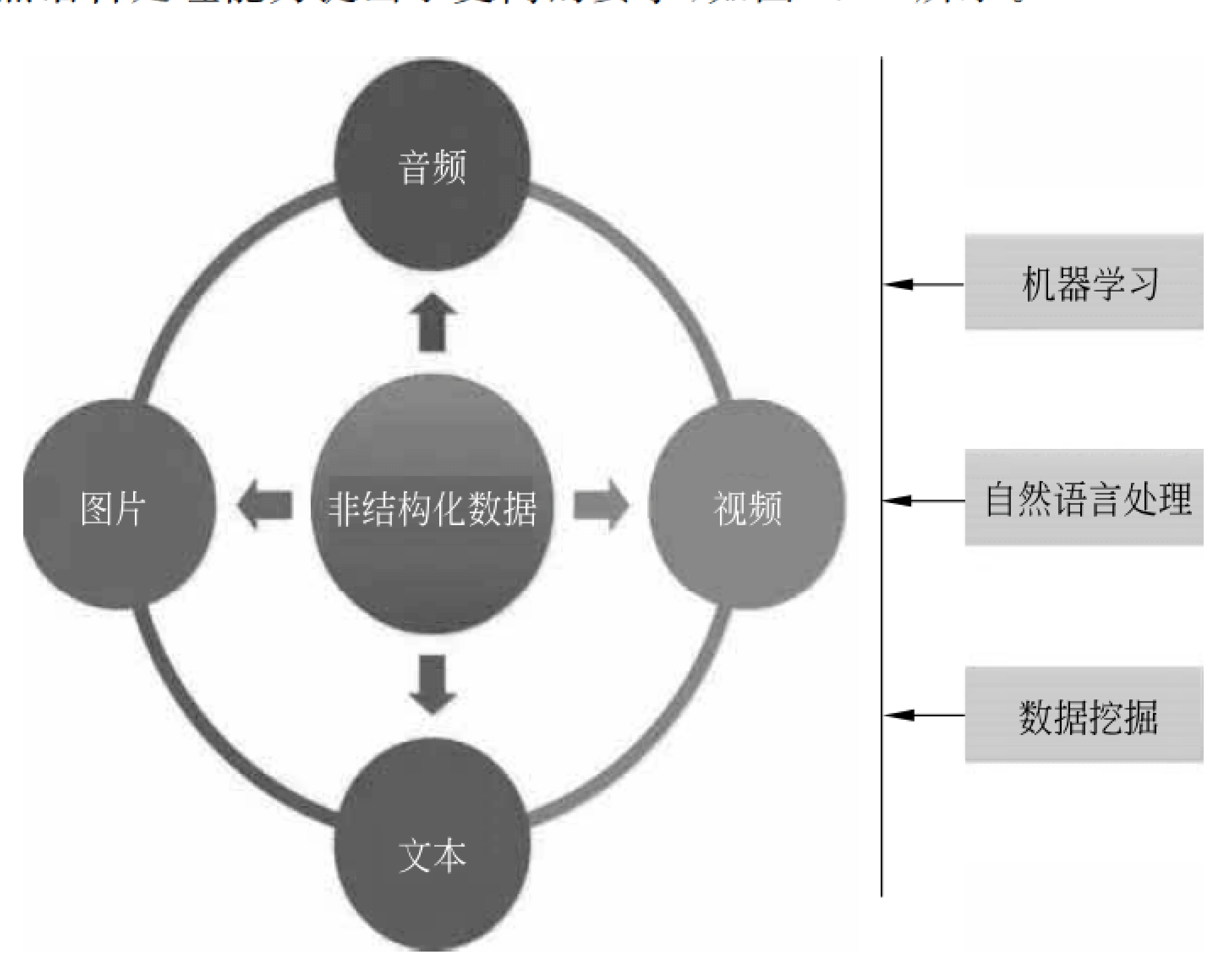


图 6.32 非结构化数据处理能力

3) 知识推理

推理能力是人类智能的重要特征,使得我们可以从已有的知识中发现隐含的知识。一般的推理往往需要一些规则的支持。例如"朋友"的"朋友",可以推理出"朋友"关系,"父亲"的"父亲"可以推理出"祖父"的关系。再比如张三的朋友很多也是李四的朋友,那我们可以推测张三和李四也很有可能是朋友关系。当然,这里会涉及概率的问题。当信息量特别多的时候,怎么把这些信息(side information)有效地与推理算法结合在一起才是最关键的。常用的推理算法包括基于逻辑(Logic)的推理和基于分布式表示方法

(Distributed Representation)的推理。随着深度学习在人工智能领域的地位变得越来越重要,基于分布式表示方法的推理也成为目前研究的热点。如果有兴趣,可以参考一下这方面目前的工作进展。

大数据、小样本、构建有效的生态闭环是关键:虽然现在能获取的数据量非常庞大,我们仍然面临着小样本问题,也就是样本数量少。假设需要搭建一个基于机器学习的反欺诈评分系统,那么首先需要一些欺诈样本。但实际上,我们能拿到的欺诈样本数量不多,即便有几百万个贷款申请,最后被标记为欺诈的样本很可能也就几万个。这对机器学习的建模提出了更高的挑战。每一个欺诈样本都是以很高昂的“代价”得到的。随着时间的推移,我们必然会收集到更多的样本,但样本的增长空间还是有局限的。这有区别于传统的机器学习系统,比如图像识别,不难拿到好几十万甚至几百万的样本。

在这种小样本条件下,构建有效的生态闭环尤其的重要。所谓的生态闭环,指的是构建有效的自反馈系统使其能够实时地反馈给我们的模型,并使得模型不断地自优化从而提升准确率。为了搭建这种自学习系统,我们不仅要完善已有的数据流系统,而且要深入到各个业务线,并对相应的流程进行优化。这也是整个反欺诈环节必要的过程,我们要知道整个过程都充满着博弈。所以需要不断地通过反馈信号来调整策略。

6.8　大数据应用案例之:数据告诉你,上海的房子都被谁买走了

事情是这样的——某年月日,学姐过来找我说:“小团啊,最近股市风起云涌变幻莫测,我觉得还是投资固定资产比较靠谱。可是,我一个外地女生在上海买得起房吗?”

我说:“学姐你收入多少?我帮你算算吧。”

学姐说:“这也太隐私啦,可不能随便告诉你,你就从整体上看一看吧。”

好吧。为了满足学姐这个毫无诚意的无理要求,我只好找出某房地产代理商提供的2014.7—2015.6上海一手房交易的抽样数据,样本数大约1万个,数据字段包括房屋价格和区位信息、购房者性别及脱敏后的身份证号(不包括姓名和末4位)等。

既然不掌握学姐的个人收入数据,那么我们只能从统计的角度看看:上海的房子都被谁买走了呢?

我们就从购房者的户籍来源、性别、星座、年龄四个角度分析一下吧。

Part1:购房者来源:上海人 VS 新上海人

我们将身份证号以“310”开头的购房者定义为“土生土长的上海人”,简称“上海人”;将其他购房者,也就是原户籍不在上海、已在上海购房的人定义为“新上海人”。

从最近一年的数据来看,购房者中上海人占比为48.5%,低于新上海人的51.5%。也就是说,上海有一半的房子被原籍意义上的“外地人”买走了。那么,新上海人都来自哪里呢?

可以看到,各省在沪购房者人数呈现明显的以上海为中心向外递减的圈层结构,即距离上海越近的地区,来沪购房者越多。

按地域片区来看，在沪购房者人数呈现出“华东＞华中＞东北＞华北＞西北＞西南＞华南”的规律。而在华东地区，原籍江苏、安徽和浙江的购房者占据了新上海人总数的41.7%。

很明显，来沪买房子的新上海人大多来自于上海周边的城市。但问题是：

是不是来自于这些地方的新上海人更热衷于买上海的房子呢？

为了回答这个问题，我们定义了各省购房者的上海买房指标 I：

I_i＝一年中在上海购房的原籍在省 i 的人数量(人)/上海外来人口中来源地为省 i 的人口数量(万人)

买房比例最高的居然是来自东北、华北和新疆的人！而在买房人数上占优的华东，买房比例反而是偏低的。总体来看：

新上海人买房比例排列最高 top3：天津、辽宁、内蒙古。

新上海人买房比例排列最低 bottom3：安徽、四川、贵州。

我想，大概北方离上海挺远，因此只有实力强大、内心坚定的北方人才会来上海发展，而且来就抱着“扎根”的信念；与之相比，从华东来上海的人数量更多、目的更多元、经济实力和个人能力差异也比较大，因此拉低了本省人在上海购房的比例。

Part2：购房者性别：男性 VS 女性

从总体来看：

最近一年的上海购房者中性别比为 147∶100；

购房者中，上海人性别比为 144∶100；

购房者中，新上海人性别比为 151∶100。

显而易见，上海的房子更多都被男性买走了。

可以看到，来自全国大部分地区的购房者都以男性居多，在沿海地区更甚。

上海购房者性别比最高原籍省 top3：广东、山东、江苏。

上海购房者性别比最低原籍省 top3：新疆、海南、宁夏。

那么，男性买房比例是不是比女性更高呢？

还是用 Part1 中定义的购房指标，我们将购房性别比与总人口性别比进行比对，计算得到新上海人中男女购房指标分别为 8.9 和 5.0。

没错，就上海而言，男性买房的比例也远比女性更高。

那么，这一差异有没有地域特征呢？

可以看到，全国大部分地区的男性在上海购房的比例都高于女性，且东部比西部差异更大。

新上海人买房男性指标最高 top3：天津、辽宁、内蒙古。

新上海人买房女性指标最高 top3：北京、宁夏、河北。

看来买房子始终还是大部分男性的核心人生任务啊。

Part3：购房者星座

接下来，我们又非常八卦地统计了最近一年在沪购房者的星座。各星座在沪购房人数如图 6.33 所示。

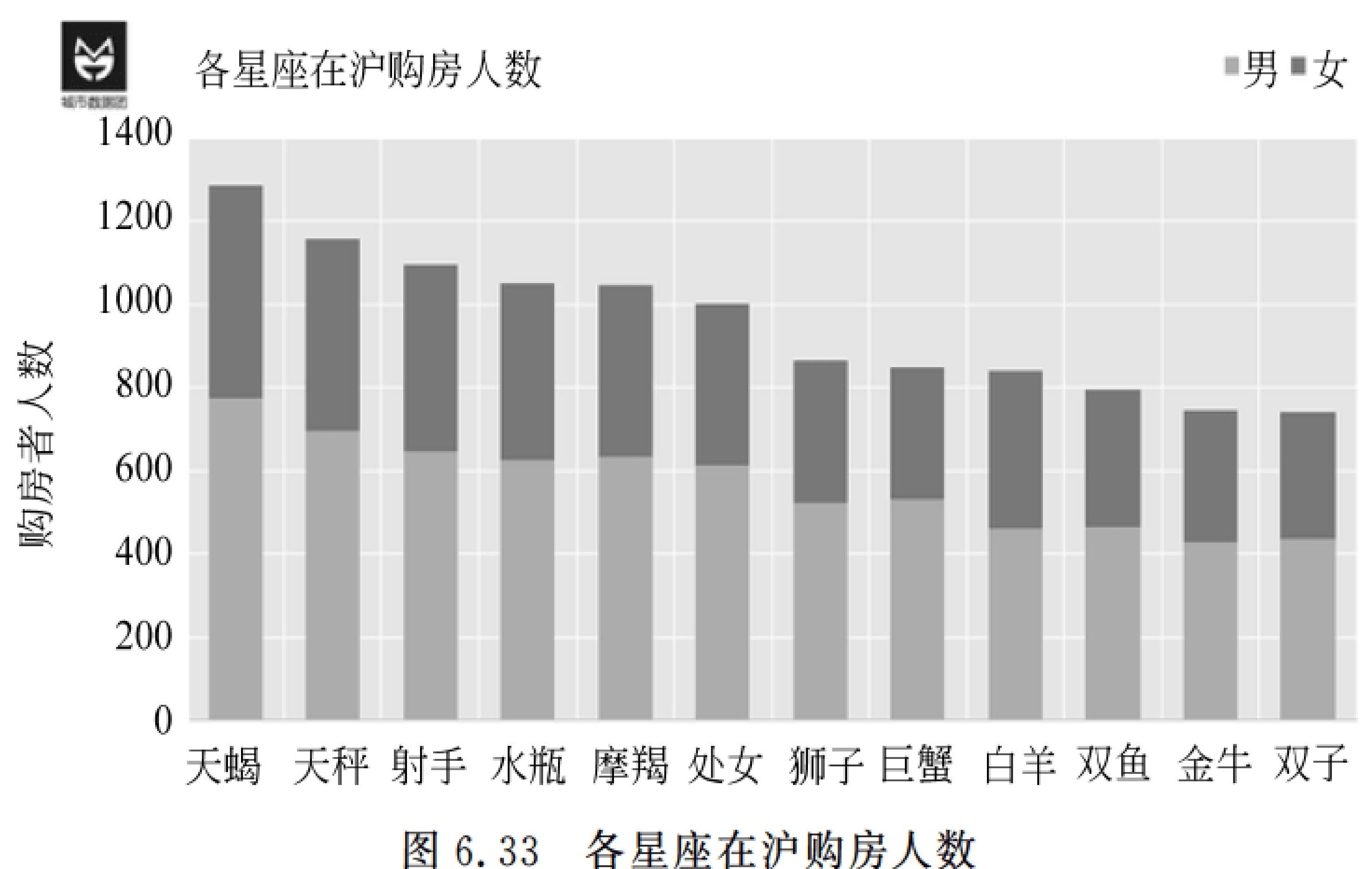


图 6.33　各星座在沪购房人数

可以看到，无论男女，天蝎、天秤和射手都稳居前三甲。

难道说，腹黑、优雅、热情可以大大提高购房成功概率？等等，这三个星座从出生日期上不是连着的吗？我好像知道了什么……

Part4：购房者年龄

我们算了一下：

上海人的购房年龄平均数为 38～39 岁；

新上海人的购房年龄平均数为 35～36 岁。

也就是说，新上海人购房比上海人要早三年（注：未区分首套房和换房）。但如果把购房者分为上海男、上海女、新上海男、新上海女四个组，并按空间圈层比较的话，会看到差异更加清晰。各圈层购房者年龄分布如图 6.34 所示。

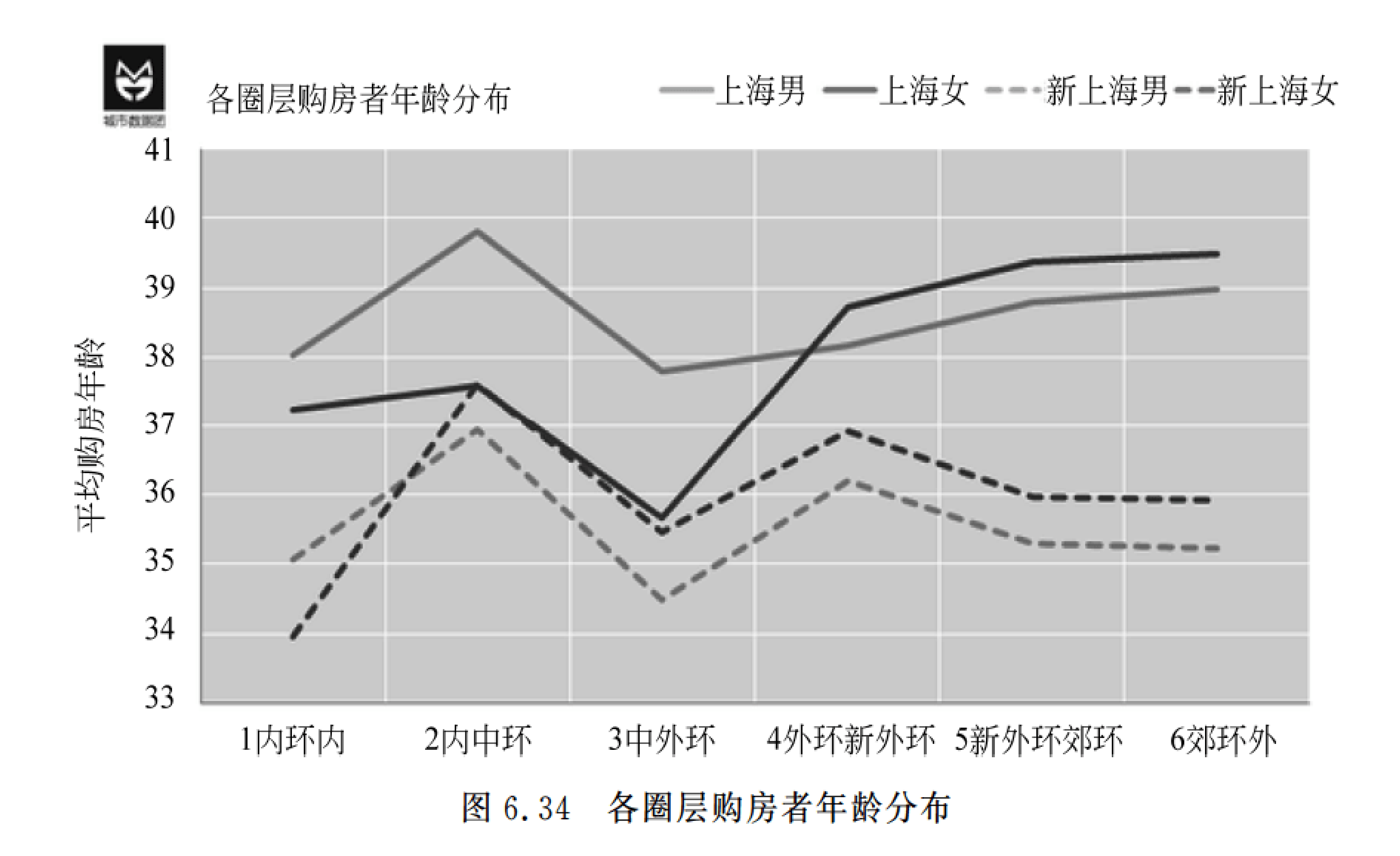


图 6.34　各圈层购房者年龄分布

可以看到：

上海男和新上海男的年龄随空间圈层的变化趋势相同，且 3 岁的年龄差异稳定存在。

但值得注意的是，市中心女性购房者年龄比男性要小，而郊区女性购房者年龄比男性要大。

接下来的问题是：什么是“好房子”呢？一千个人心中有一千个哈姆雷特。为了回答这个问题，我们不妨简单粗暴地认为市中心的就是好房子。

我们仍然按照四组人购买的房子的区位进行统计，如图 6.35 所示。

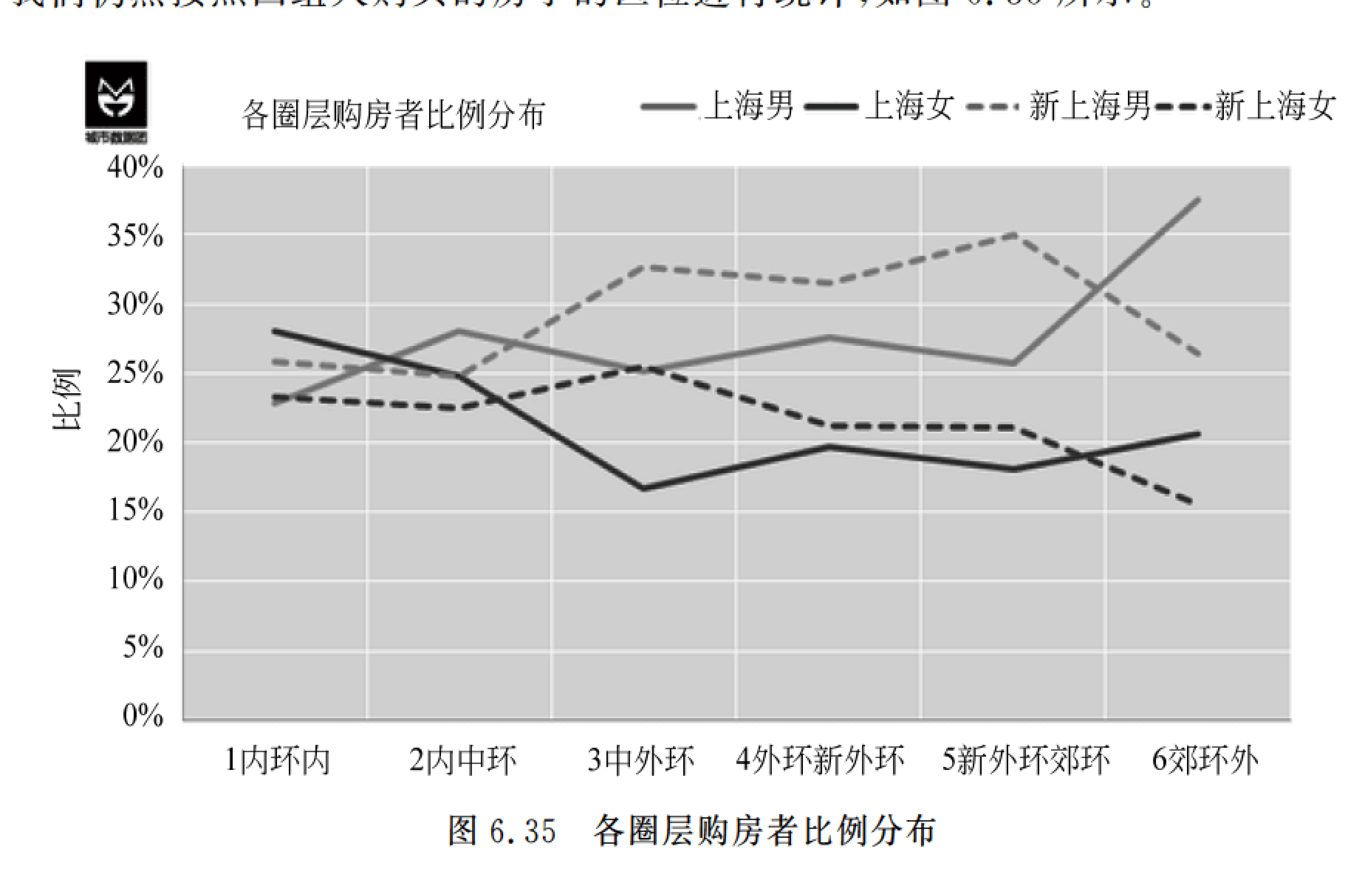


图 6.35 各圈层购房者比例分布

如图 6.36 可知：

市中心(内环以内)，上海女＞新上海男＞新上海女＞上海男；

中心城区(外环以内)，新上海男＞上海男＞新上海女＞上海女。

简单地说，就是上海中心城区的新上海人比上海人更多，更多的好房子被新上海人买走了。

这是为什么呢？我猜可能是由于以下原因：

从外地来到上海发展，并买房成为新上海人的，本身就拥有较强的个人能力或经济实力。

上海人只能在上海买房，个人能力和经济实力参差不齐，因此在市中心和郊区都会买房(去其他地方发展的上海人数量很少，忽略不计)。

为了印证这个猜想，我又用了新上海人购房的总价与其原籍省的人均 GDP 进行了比较，如图 6.36 所示。

如图 6.36 可知，二者间的正相关的关系还是比较明显的。也就是说，买什么样的房，跟地区和家庭的经济实力有着很大的关系。

再对性别进行比较的话，我们会发现：从市中心向郊区，购房者性别比呈增加趋势，也就是说女性买房比男性更靠近市中心。这一点在新上海人中更为显著。

各圈层购屋者性别比如图 6.37 所示。

数据来源说明：

(1) 房屋销售和购房者数据来源于同策房产咨询。

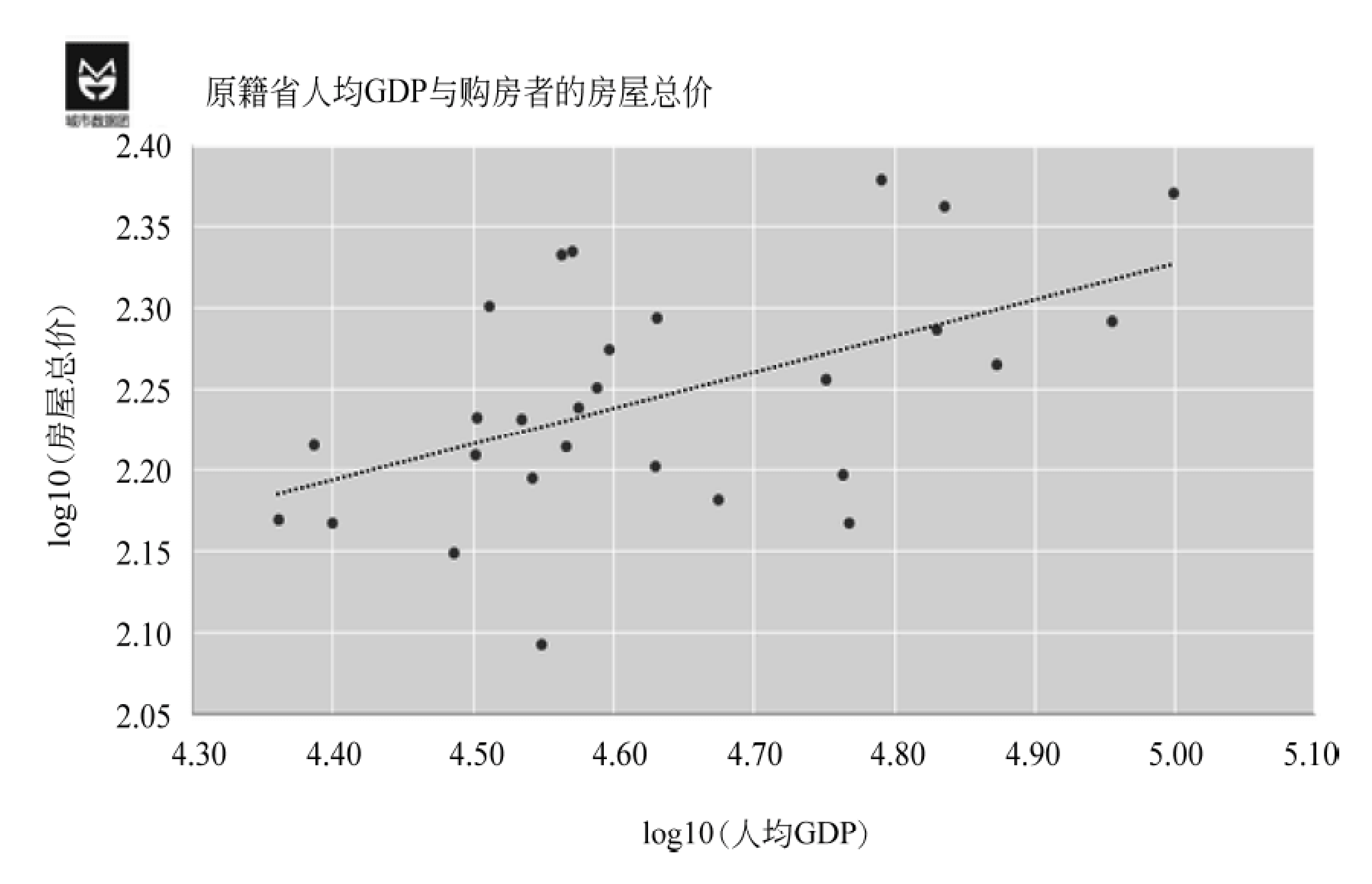


图 6.36　原籍省人均 GDP 与购房者的房屋总价

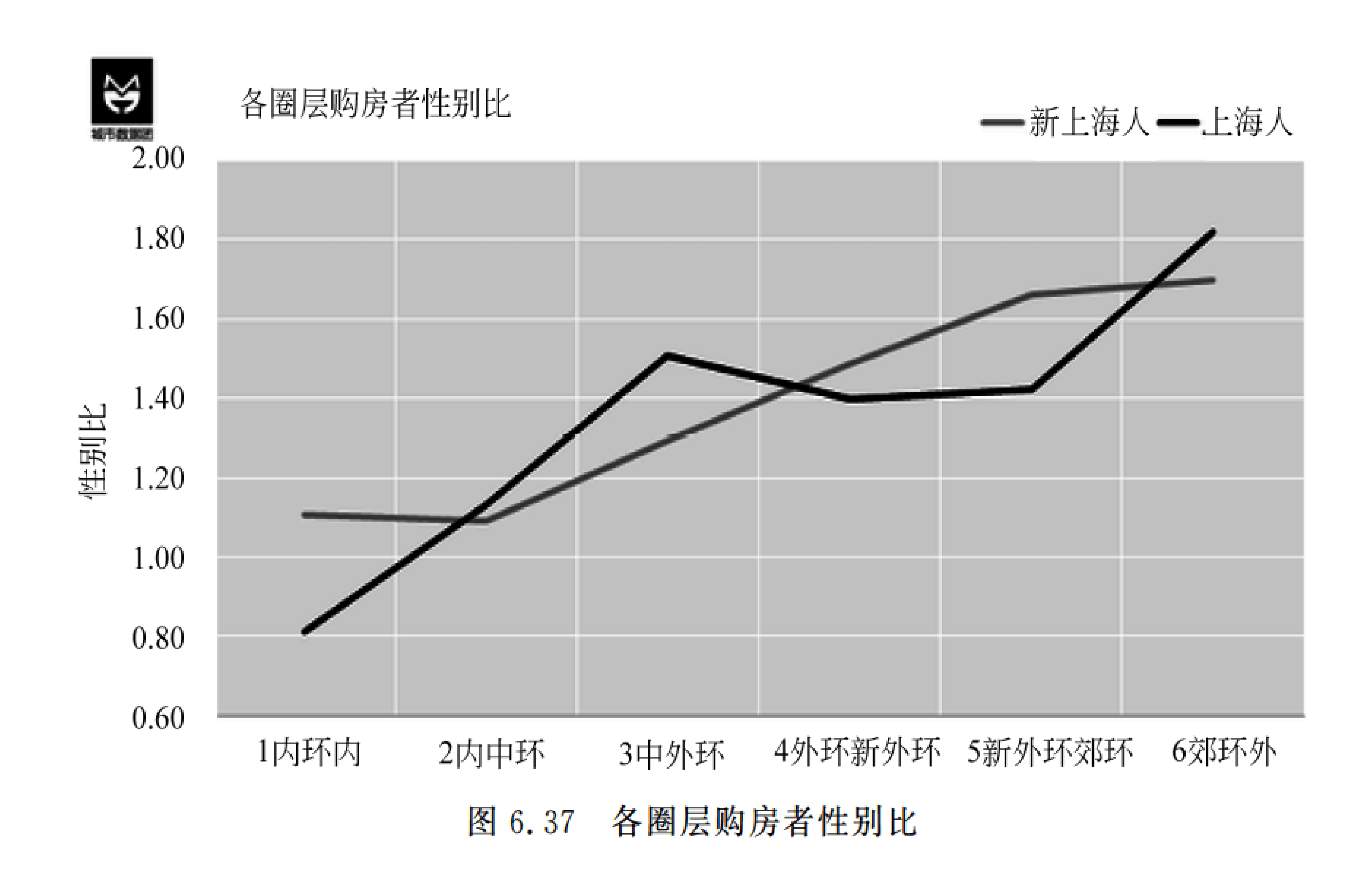


图 6.37　各圈层购房者性别比

(2) 其他数据来源于 2010 年上海市人口普查数据、上海统计年鉴 2014 等。

习题与思考题

一、选择题

1. 下列哪一项不是大数据提供的用户交互方式？(　　)

   A. 统计分析和数据挖掘　　　　B. 任意查询和分析

   C. 图形化展示　　　　　　　　D. 企业报表

2. 关于大数据和互联网，以下哪些说法是正确的？(　　)(多选题)

   A. 互联网的出现使得监视变得更容易、成本更低廉也更有用处

B. 大数据不管如何运用都是我们合理决策过程中的有力武器

C. 大数据的价值不再单纯来源于它的基本用途，而更多源于它的二次利用

D. 大数据时代，很多数据在收集的时候并无意用作其他用途，而最终却产生了很多创新性的用途

3. 在网络爬虫的爬行策略中，应用最为基础的是(　　)。(多选题)

A. 深度优先遍历策略　　B. 广度优先遍历策略

C. 高度优先遍历策略　　D. 反向链接策略

E. 大站优先策略

4. 大数据科学关注大数据网络发展和运营过程中(　　)大数据的规律及其与自然和社会活动之间的关系。

A. 大数据网络发展和运营过程　　B. 规划建设运营管理

C. 规律和验证　　D. 发现和验证

5. 大数据的价值是通过数据共享、(　　)后获取最大的数据价值。

A. 算法共享　　B. 共享应用　　C. 数据交换　　D. 交叉复用

6. IBM 大数据平台和应用程序框架，(　　)以经济高效的方式分析 PB 级的结构化和非结构化信息。

A. 流计算　　B. Hadoop　　C. 数据仓库　　D. 语境搜索

7. 临床决策支持系统通过电子病历、医学指导的比较等提高手术质量，降低错误治疗和(　　)。

A. 医疗事故　　B. 病患投诉　　C. 民事诉讼　　D. 手术费用

8.《数据新闻学手册》的作者们认为，通过数据的使用，记者工作的重点从“第一个报道者”转化成为对特定事件的影响的(　　)。

A. 拍摄者　　B. 知情者　　C. 记录者　　D. 阐释者

9. 通过(　　)和展示数据背后的(　　)，运用丰富的、具有互动性的可视化手段，数据新闻学成为新闻学作为一门新的分支进入主流媒体，即用数据报道新闻。

A. 数据收集　　B. 数据挖掘　　C. 真相　　D. 关联与模式

10. 什么是 KDD?(　　)

A. 数据挖掘与知识发现　　B. 领域知识发现

C. 文档知识发现　　D. 动态知识发现

二、问答题

1. 简述数据库与信息检索技术的比较。

2. 解释 WEB 搜索引擎工作原理。

3. 大数据索引和查询是如何进行的?

4. 概述数据可视化定义与应用。

5. 概述知识图谱的概念和应用。

第7章　大数据分析与数据挖掘

7.1　大数据的分析及应用

7.1.1　数据处理和分析的发展

1. 传统方式的数据处理和分析

传统上，为了特定分析目的进行的数据处理都是基于相当静态的蓝图。通过常规的业务流程，企业通过 CRM、ERP 和财务系统等应用程序，创建基于稳定数据模型的结构化数据。数据集成工具用于从企业应用程序和事务型数据库中提取、转换和加载数据到一个临时区域，在这个临时区域进行数据质量检查和数据标准化，数据最终被模式化到整齐的行和表。这种模型化和清洗过的数据被加载到企业级数据仓库。这个过程会周期性发生，如每天或每周，有时会更频繁。数据处理分析资料的流程如图 7.1 所示。

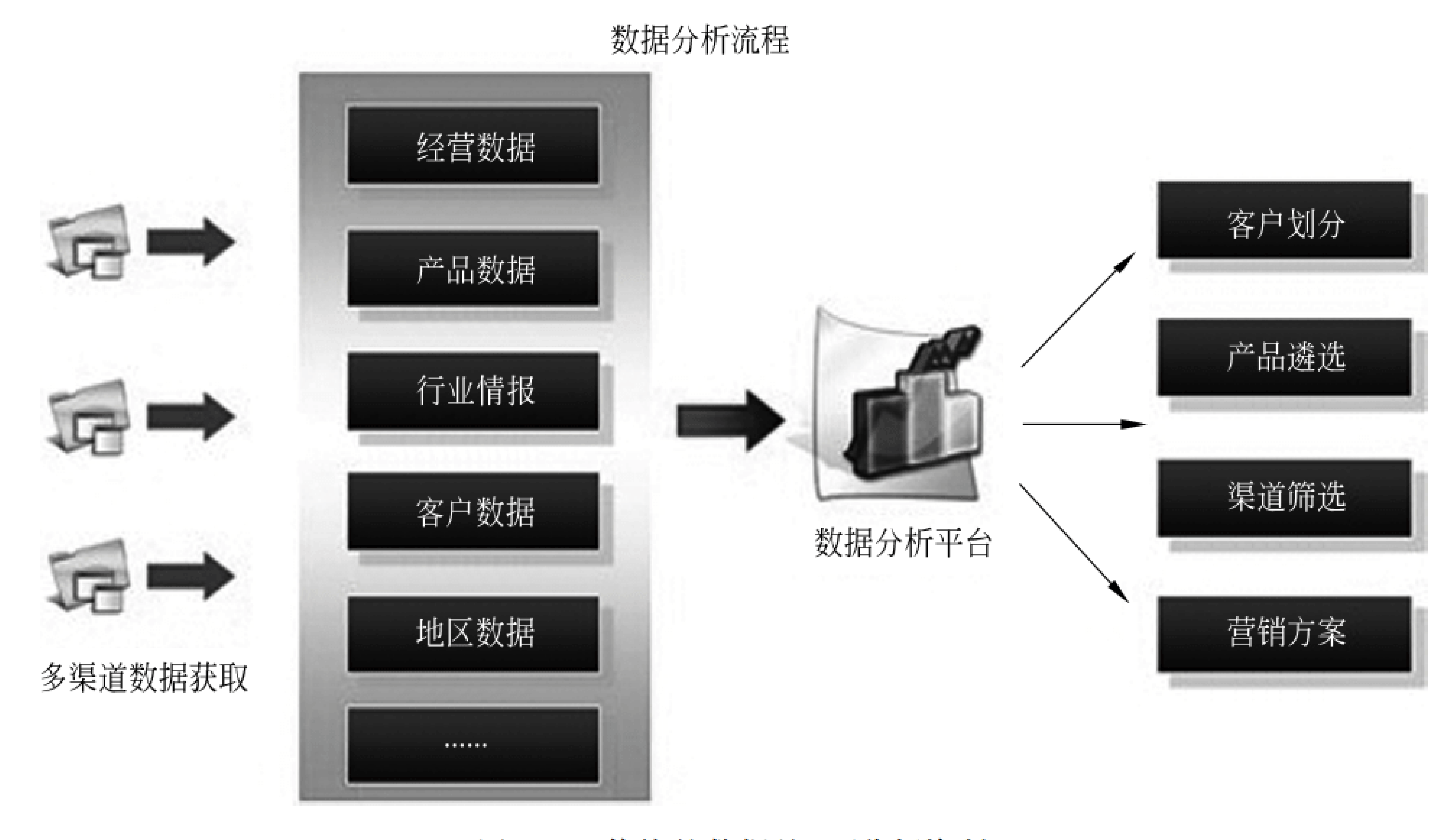


图 7.1　传统的数据处理/分析资料

在传统数据仓库中，数据仓库管理员创建计划，定期计算仓库中的标准化数据，并将产生的报告分配到各业务部门。他们还为管理人员创建仪表板和其他功能有限的可视化工具。

同时，业务分析师利用数据分析工具在数据仓库进行高级分析，或者通常情况下，由于数据量的限制，将样本数据导入到本地数据库中。非专业用户通过前端的商业智能工

具(SAP 的 BusinessObjects 和 IBM 的 Cognos)对数据仓库进行基础的数据可视化和有限的分析。传统数据仓库的数据量很少超过几 TB,因为大容量的数据会占用数据仓库资源并且降低性能。

2. 大数据处理和分析的新方法

存在多种方法处理和分析大数据,但多数都有一些共同的特点。即利用硬件的优势,使用扩展的、并行的处理技术,采用非关系型数据存储处理非结构化和半结构化数据,并对大数据运用高级分析和数据可视化技术,向终端用户传达见解。

毋庸置疑,现在大数据平台和大数据分析工具日益普及,作用是可以帮助企业收集和分析数据,好处是可以寻找有价值的商业信息和洞察,以改进产品与服务。大数据分析工具用于分析数据,可以开发预测模型(predictive model)和规范模型(prescriptive model)。在现代化的业务流程应用中,嵌入这些模型能够提高企业的生产力和价值。同时,使用大数据分析工具可以轻松进行扩展,获取通常在大数据平台才有的可用资源。

其实,大数据分析工具经常提供的技术,一般而言,都不算什么新鲜事物。只是到最近这几年,数据挖掘算法的强大功能才被主流商业用户采用,它可以结合海量数据、多种数据类型和不同的数据结构,对数据集进行预测性分析(predictive analyses)和规范性分析(prescriptive analyses)。

但在用户看来,大数据分析仍然是一种新兴的企业级功能,要想靠它达到预期收益,一定存在风险,还要投入很大的时间成本。所以,在决定投身之前,一定要弄清楚怎样判断什么样的大数据分析适合你的企业?

7.1.2 大数据分析面对的数据类型

有一个概念可以很清楚地区分大数据分析和其他形式的分析:要分析的数据有多大的数据量?数据规模如何?数据是否呈多样性?在过去,通常是从非常大的数据库中提取样本数据集,建立分析模型,然后通过测试再调整的过程加以改进。而现在,随着计算平台能够提供可扩展的存储和计算能力,可分析的数据量几乎不再受任何限制。这意味着,实时预测性分析和访问大量正确的数据可以帮助企业改善业绩。这样的机会取决于企业能否整合和分析不同类型大数据。以下四大类数据就是大数据要分析的数据类型。

1. 交易数据(Transaction data)

大数据平台能够获取时间跨度更大、更海量的结构化交易数据,这样就可以对更广泛的交易数据类型进行分析,不仅仅包括 POS 或电子商务购物数据,还包括行为交易数据,例如 Web 服务器记录的互联网点击流数据日志。

2. 人为数据(Human-generated data)

非结构数据广泛存在于电子邮件、文档、图片、音频、视频,以及通过博客、维基,尤其是社交媒体产生的数据流。这些数据为使用文本分析功能进行分析提供了丰富的数据源。

3. 移动数据(Mobile data)

能够上网的智能手机和平板越来越普遍。这些移动设备上的 App 都能够追踪和沟

通无数事件,从 App 内的交易数据(如搜索产品的记录事件)到个人信息资料或状态报告事件(如地点变更即报告一个新的地理编码)。

4. 机器和传感器数据(Machine and sensor data)

机器和传感器数据包括功能设备创建或生成的数据,例如智能电表、智能温度控制器、工厂机器和连接互联网的家用电器生成的数据。这些设备可以配置为与互联网络中的其他结点通信,还可以自动向中央服务器传输数据,这样就可以对数据进行分析。机器和传感器数据是来自新兴的物联网(IoT)所产生的主要例子。来自物联网的数据可以用于构建分析模型,连续监测预测性行为(如当传感器值表示有问题时进行识别),提供规定的指令(如警示技术人员在真正出问题之前检查设备)。

7.1.3 大数据分析与处理方法

越来越多的应用涉及大数据,这些大数据的属性,包括数量、速度、多样性等等都呈现了大数据不断增长的复杂性,所以,大数据的分析方法在大数据领域就显得尤为重要,可以说是判定最终信息是否有价值的决定性因素。基于此,大数据分析的方法理论有哪些呢?

1. 大数据分析的五个基本方面

1) 预测性分析能力(Predictive Analytic Capabilities)

数据挖掘可以让分析员更好的理解数据,而预测性分析可以让分析员根据可视化分析和数据挖掘的结果做出一些预测性的判断。

2) 数据质量和数据管理(Data Quality and Master Data Management)

数据质量和数据管理是一些管理方面的最佳实践。通过标准化的流程和工具对数据进行处理可以保证一个预先定义好的高质量的分析结果。

3) 可视化分析(Analytic Visualizations)

不管是对数据分析专家还是普通用户,数据可视化都是数据分析工具最基本的要求。可视化可以直观地展示数据,让数据自己说话,让观众听到结果。

4) 语义引擎(Semantic Engines)

我们知道由于非结构化数据的多样性带来了数据分析的新的挑战,我们需要一系列的工具去解析、提取、分析数据。语义引擎需要被设计成能够从“文档”中智能提取信息。

5) 数据挖掘算法(Data Mining Algorithms)

可视化是给人看的,数据挖掘就是给机器看的。集群、分割、孤立点分析还有其他的算法让我们深入数据内部,挖掘价值。这些算法不仅要处理大数据的量,也要处理大数据的速度。

假如大数据真的是下一个重要的技术革新,我们最好把精力关注在大数据能给我们带来的好处,而不仅仅是挑战。

7.1.4 数据分析的步骤

什么是数据分析?数据分析是用适当的统计分析方法对收集来的大量数据进行分析,将它们加以汇理解并消化,以求最大化地开发数据的功能,发挥数据的作用。数据分

析的目的？把隐藏在一大批看似杂乱无章的数据背后的信息集中和提炼出来，总结出研究对象的内在规律。

1. 数据分析的目的

把隐藏在一大批看似杂乱无章的数据背后的信息集中和提炼出来，总结出研究对象的内在规律。

2. 数据分析的分类

数据分析主要有三大作用：现状分析、原因分析、预测分析，分别反映了数据分析的描述性、探索性和验证性，如图 7.2 所示。

数据分析
描述性 探索性 验证性

图 7.2 数据分析的三大作用

3. 数据分析的六部曲

数据分析流程主要分为六个步骤，如图 7.3 所示。

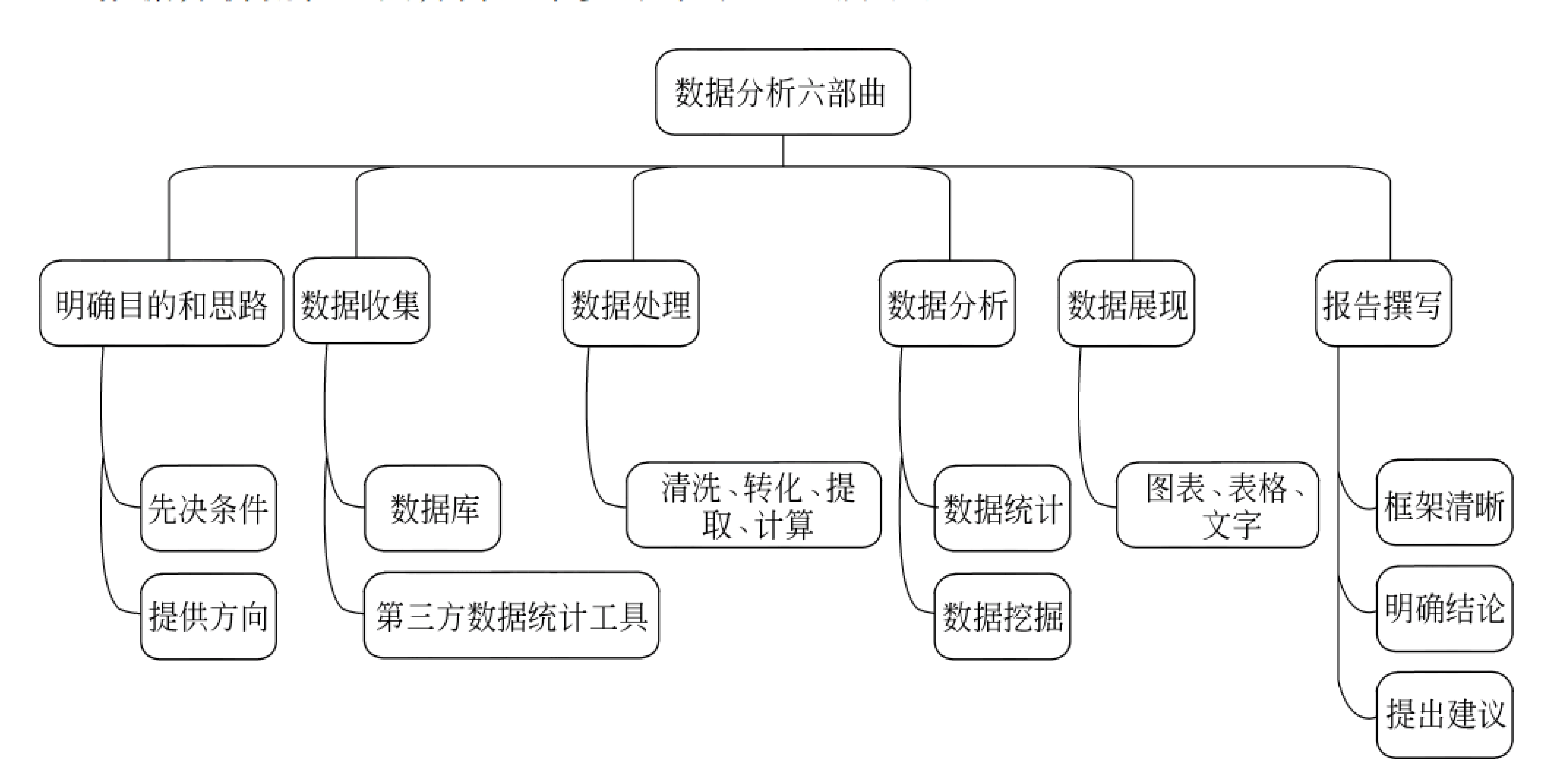


图 7.3 数据分析流程的六个步骤

1）明确目的和思路

梳理分析思路，并搭建分析框架，把分析目的分解成若干个不同的分析要点，即如何具体开展数据分析，需要从哪几个角度进行分析，采用哪些分析指标(各类分析指标需合理搭配使用)。同时，确保分析框架的体系化和逻辑性。

2）数据收集

一般数据来源于四种方式：数据库、第三方数据统计工具、专业的调研机构的统计年鉴或报告(如艾瑞资讯)、市场调查。

对于数据的收集需要预先做埋点，在发布前一定要经过谨慎的校验和测试，因为一旦版本发布出去而数据采集出了问题，就获取不到所需要的数据，影响分析效果。

3）数据处理

数据处理主要包括数据清洗、数据转化、数据提取、数据计算等处理方法，将各种原始数据加工成为产品经理需要的直观的可看数据。

4）数据分析

数据分析是用适当的分析方法及工具，对处理过的数据进行分析，提取有价值的信息，形成有效结论的过程。

常用的数据分析工具，掌握 Excel 的数据透视表，就能解决大多数的问题。需要的话，可以再有针对性的学习 SPSS、SAS 等。

数据挖掘是一种高级的数据分析方法，侧重解决四类数据分析问题：分类、聚类、关联和预测，重点在寻找模式与规律。

5）数据展现

一般情况下，数据是通过表格和图形的方式来呈现的。常用的数据图表包括饼图、柱形图、条形图、折线图、气泡图、散点图、雷达图等。进一步加工整理变成我们需要的图形，如金字塔图、矩阵图、漏斗图、帕雷托图等。数据展现的图表如图 7.4 所示。

要表达的数据和信息	饼图	柱形图	条形图	折线图	气泡图	其他
成分（整体的一部分）						
排序（数据的比较）						
时间序列（走势、趋势）						
频率分布（数据频次）						
相关性（数据的关系）						
多重数据比较						

图 7.4　数据展现的图表

一般能用图说明问题的就不用表格，能用表说明问题的就不用文字。

图表制作的五个步骤如下：

（1）确定要表达主题；

（2）确定哪种图表最适合；

（3）选择数据制作图表；

（4）检查是否真实反映数据；

（5）检查是否表达观点。

6）报告撰写

一份好的数据分析报告，首先需要有一个好的分析框架，并且图文并茂、层次明晰，能够让阅读者一目了然。结构清晰、主次分明可以使阅读者正确理解报告内容；图文并茂，可以令数据更加生动活泼，提高视觉冲击力，有助于阅读者更形象、直观地看清楚问题和结论，从而产生思考。

好的数据分析报告需要有明确的结论、建议或解决方案。

4. 数据分析的四大误区

(1) 分析目的不明确,为了分析而分析。

(2) 缺乏行业、公司业务认知,分析结果偏离实际。数据必须和业务结合才有意义。摸清楚所在产业链的整个结构,对行业的上游和下游的经营情况有大致的了解,再根据业务当前的需要,制定发展计划,归类出需要整理的数据。同时,熟悉业务才能看到数据背后隐藏的信息。

(3) 为了方法而方法,为了工具而工具,只要能解决问题的方法和工具就是好的方法和工具。

(4) 数据本身是客观的,但被解读出来的数据是主观的。同样的数据由不同的人分析很可能得出完全相反的结论,所以一定不能提前带着观点去分析。

7.1.5 大数据分析应用

1. 大数据分析应用场景

假如以下应用场景听上去那么像你所在的企业,你可要认真开始考虑大数据分析工具,这将是一项合理的投资!

1) 客户分析(Customer analytics)

这包括分析客户的信息资料、行为和特点到开发模型,对客户进行细分、预测流失以及提供帮助挽留客户的下一个最好报价。

2) 营销分析(Sales and marketing analytics)

有两种营销用例。第一种是使用营销模型,改进面向客户的应用程序,更好地向客户提供推荐。例如,更好地识别交叉销售和追加销售机会,减少放弃的购物车,总体提升集成推荐引擎的准确性。第二种更具反思性,因为它是为了展示营销部门过程和活动的表现,并建议进行调整,以优化绩效。例如,分析哪个活动解决了确认群体的需求,或激励活动付诸行动的成功率。

3) 社交媒体分析(Social media analytics)

通过不同社交媒体渠道生成的内容为分析客户情感和舆情监督提供了丰富的资料。

4) 网络安全(Cyber security)

大规模网络安全事件(如对美国零售商 Target、Sony 的网络攻击)的发生,让企业越来越意识到网络攻击发生时快速识别的重要性。识别潜在的攻击包括建立分析模型,监测大量网络活动数据和相应的访问行为,以识别可能进行入侵的可疑模式。

5) 设备管理(Plant and facility management)

随着越来越多的设备和机器能够与互联网相连,企业能够收集和分析传感器数据流,包括连续用电、温度、湿度和污染物颗粒等无数潜在变量。模型还可以预测设备故障,安排预防性的维护,以确保项目正常进行,不中断。

6) 管道管理(Pipeline management)

越来越多的能源管道具有传感器和通信功能。连续的传感器数据可以用来分析本地

和全球性问题，表示是否需要引起注意或进行维护。

7）供应链和渠道分析(Supply chain and channel analytics)

通过对仓库库存、POS 交易和多种渠道的运输（如陆运、铁路、海运）进行分析，可建立预测分析模型，有效帮助预先补货，制定库存管理策略，管理物流，以及因延迟危及及时交货时对线路进行优化并发送通知。

8）价格优化(Price optimization)

零售商希望最大限度地提高产品销售的整体盈利，建立的分析模型可以结合不同种类的数据流，包括竞争对手的价格、跨不同地域的销售交易数据(以查看需求)，以及生产、库存和供应链的信息(以监测供货)。这样的模型可以动态地调整产品价格：当供不应求时，或竞争对手没货时，价格上涨；当因季节变化需清理库存时，价格下调。

9）欺诈行为检测(Fraud detection)

身份盗用事件不断增长，随之而来的是欺诈行为和交易的不断增长。金融机构对上亿条的交易数据进行分析，以识别欺诈行为模式。这样的分析模型还可以在潜在欺诈交易可能发生时，向用户发送警示。

所有这些应用场景都具有相似的特点，即分析涉及结构化和非结构化数据，被访问的数据或数据流来自不同来源，以及数据量可能巨大。反之，对数据进行分析可以建立分析模型，用于实时识别来自同一数据源和数据流的模式。

2. 大数据分析技术

让 Hadoop 和其他大数据技术如此引人注目的部分原因是，它们让企业找到问题的答案，而在此之前企业甚至不知道问题是什么。这可能会产生引出新产品的想法，或者帮助确定改善运营效率的方法。不过，也有一些已经明确的大数据用例，无论是互联网巨头如 Google、Facebook 和 LinkedIn，还是更多的传统企业。

1）推荐引擎

网络资源和在线零售商使用 Hadoop 根据用户的个人资料和行为数据匹配和推荐用户、产品和服务。LinkedIn 使用此方法增强其“你可能认识的人”这一功能，而亚马逊利用该方法为网上消费者推荐相关产品。

2）情感分析

Hadoop 与先进的文本分析工具结合，分析社会化媒体和社交网络发布的非结构化的文本，包括 Tweets 和 Facebook，以确定用户对特定公司、品牌或产品的情绪。分析既可以专注于宏观层面的情绪，也可以细分到个人用户的情绪。

3）风险建模

财务公司、银行等公司使用 Hadoop 和下一代数据仓库分析大量交易数据，以确定金融资产的风险，模拟市场行为为潜在的“假设”方案做准备，并根据风险为潜在客户打分。

4）欺诈检测

金融公司、零售商等使用大数据技术将客户行为与历史交易数据结合来检测欺诈行为。例如，信用卡公司使用大数据技术识别可能的被盗卡的交易行为。

5）营销活动分析

各行业的营销部门长期使用技术手段监测和确定营销活动的有效性。大数据让营销团队拥有更大量的越来越精细的数据，如点击流数据和呼叫详情记录数据，以提高分析的准确性。

6）客户流失分析

企业使用 Hadoop 和大数据技术分析客户行为数据并确定分析模型，该模型指出哪些客户最有可能流向存在竞争关系的供应商或服务商。企业就能采取最有效的措施挽留即将流失客户。

7）社交图谱分析

Hadoop 和下一代数据仓库相结合，通过挖掘社交网络数据，可以确定社交网络中哪些客户对其他客户产生最大的影响力。这有助于企业确定其“最重要”的客户，不总是那些购买最多产品或花最多钱的，而是那些最能够影响他人购买行为的客户。

8）用户体验分析

面向消费者的企业使用 Hadoop 和其他大数据技术将之前单一客户互动渠道（如呼叫中心、网上聊天、微博等）数据整合在一起，以获得对客户体验的完整视图。这使企业能够了解客户交互渠道之间的相互影响，从而优化整个客户生命周期的用户体验。

9）网络监控

Hadoop 和其他大数据技术被用来获取、分析和显示来自服务器、存储设备和其他 IT 硬件的数据，使管理员能够监视网络活动、诊断瓶颈等问题。这种类型的分析，也可应用到交通网络，当然也可以应用到其他网络。

10）研究与发展

有些企业（如制药商）使用 Hadoop 技术进行大量文本及历史数据的研究，以协助新产品的开发。

当然，上述这些都只是大数据用例的举例。事实上，在所有企业中大数据最引人注目的用例可能尚未被发现。这就是大数据的希望。

7.2 数据挖掘技术

数据挖掘（Data Mining，DM）又称数据库中的知识发现（Knowledge Discover in Database，KDD），是目前人工智能和数据库领域研究的热点问题，所谓数据挖掘，是指从数据库的大量数据中揭示出隐含的、先前未知的并有潜在价值的信息的非平凡过程。数据挖掘是一种决策支持过程，它主要基于人工智能、机器学习、模式识别、统计学、数据库、可视化技术等，高度自动化地分析企业的数据，做出归纳性的推理，从中挖掘出潜在的模式，帮助决策者调整市场策略，减少风险，做出正确的决策。

7.2.1 数据挖掘的定义

1. 技术上的定义及含义

数据挖掘（Data Mining）就是从大量的、不完全的、有噪声的、模糊的、随机的实际应

用数据中，提取隐含在其中的、人们事先不知道的但又是潜在有用的信息和知识的过程。这个定义包括好几层含义：数据源必须是真实的、大量的、含噪声的；发现的是用户感兴趣的知识；发现的知识要可接受、可理解、可运用；并不要求发现放之四海皆准的知识，仅支持特定的发现问题。

与数据挖掘相近的同义词有数据融合、人工智能、商务智能、模式识别、机器学习、知识发现、数据分析和决策支持等。

从广义上理解，数据、信息也是知识的表现形式，但是人们更把概念、规则、模式、规律和约束等看作知识。人们把数据看作是形成知识的源泉，好像从矿石中采矿或淘金一样。原始数据可以是结构化的，如关系数据库中的数据；也可以是半结构化的，如文本、图形和图像数据；甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的，也可以是非数学的；可以是演绎的，也可以是归纳的。发现的知识可以被用于信息管理、查询优化、决策支持和过程控制等，还可以用于数据自身的维护。

因此，数据挖掘是一门交叉学科，它把人们对数据的应用从低层次的简单查询，提升到从数据中挖掘知识，提供决策支持。在这种需求的牵引下，汇聚了不同领域的研究者，尤其是数据库技术、人工智能技术、数理统计、可视化技术、并行计算等方面的学者和工程技术人员，投身到数据挖掘这一新兴的研究领域，形成新的技术热点。

这里所说的知识发现，不是要求发现放之四海而皆准的真理，也不是要去发现崭新的自然科学定理和纯数学公式，更不是什么机器定理证明。实际上，所有发现的知识都是相对的，是有特定前提和约束条件，面向特定领域的，同时还要能够易于被用户理解。最好能用自然语言表达所发现的结果。

数据挖掘对知识特征的揭示如图 7.5 所示。

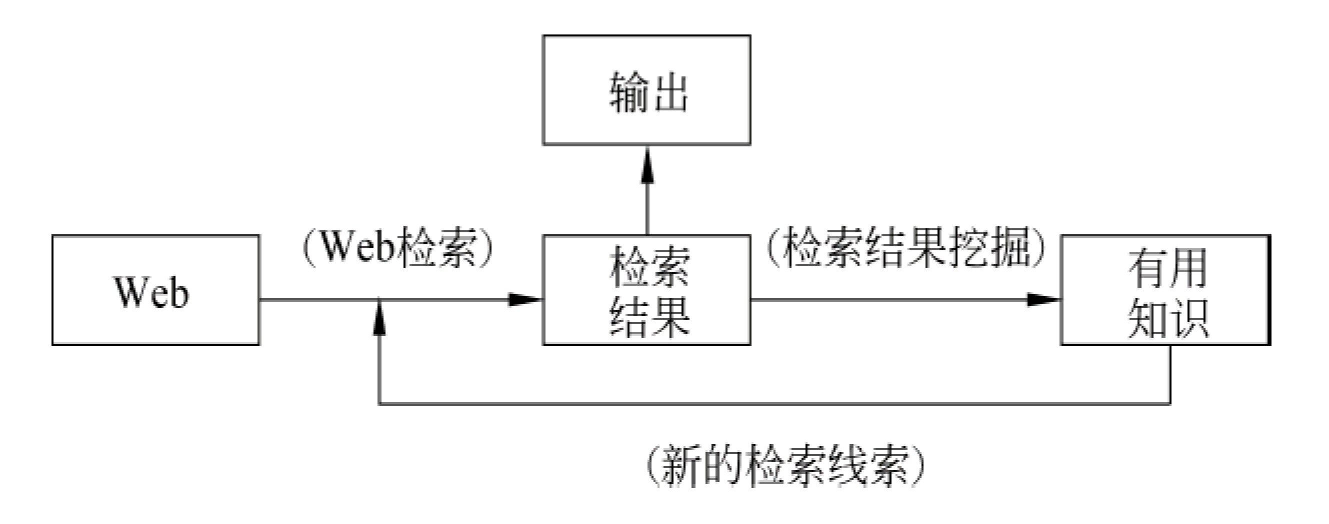


图 7.5　数据挖掘揭示知识特征

2. 商业角度的定义

数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。

简言之，数据挖掘其实是一类深层次的数据分析方法。数据分析本身已经有很多年的历史，只不过在过去数据收集和分析的目的是用于科学研究，另外，由于当时计算能力的限制，对大数据量进行分析的复杂数据分析方法受到很大限制。

现在，由于各行业业务自动化的实现，商业领域产生了大量的业务数据，这些数据不再是为了分析的目的而收集的，而是由于纯机会的(Opportunistic)商业运作而产生。分析这些数据也不再是单纯为了研究的需要，更主要是为商业决策提供真正有价值的信息，

进而获得利润。但所有企业面临的一个共同问题是：企业数据量非常大，而其中真正有价值的信息却很少，因此从大量的数据中经过深层分析，获得有利于商业运作、提高竞争力的信息，就像从矿石中淘金一样，数据挖掘也因此而得名。基于数据仓库的数据挖掘如图 7.6 所示。

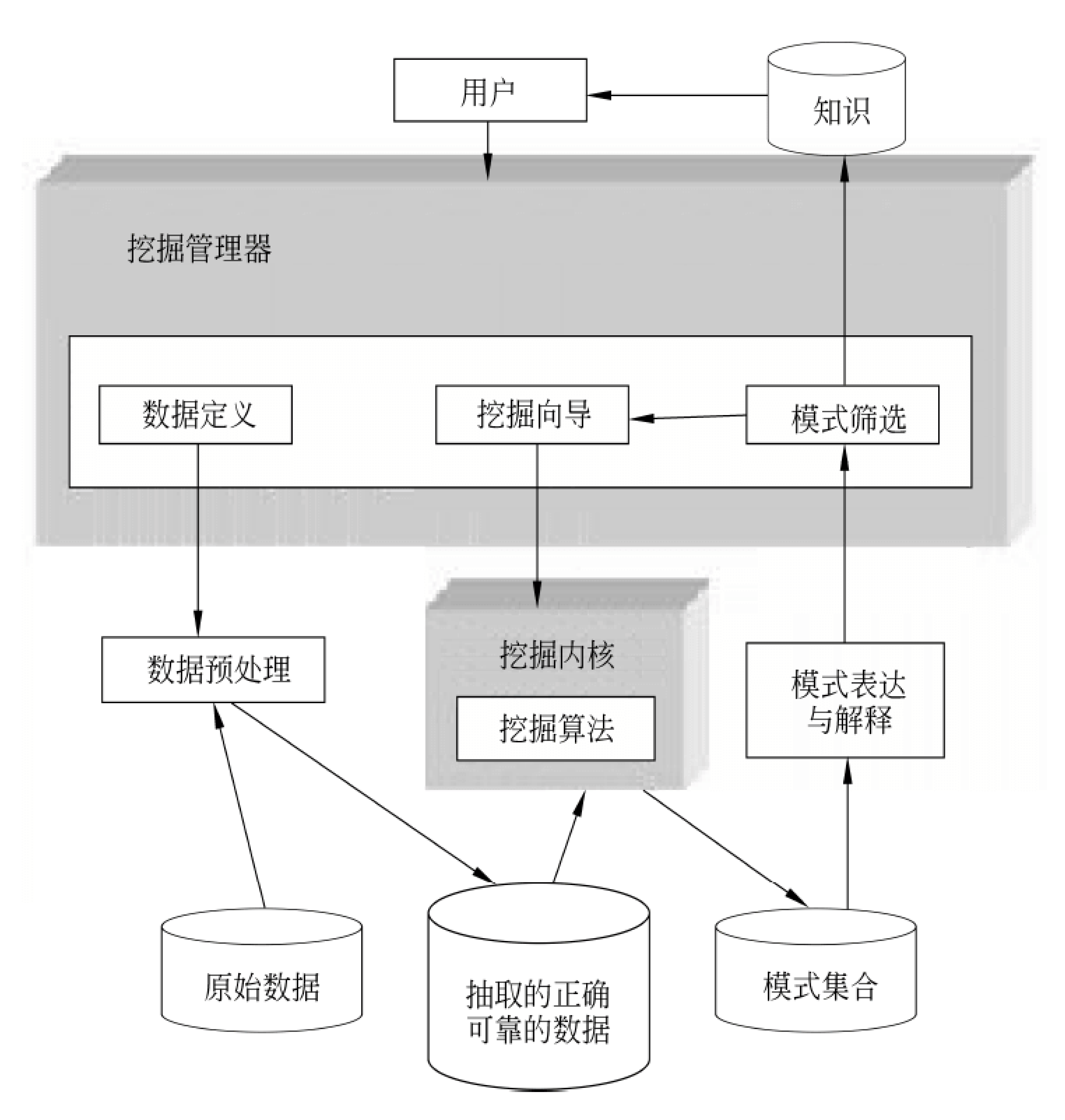


图 7.6　基于数据仓库的数据挖掘

因此，数据挖掘可以描述为：按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证已知的规律性，并进一步将其模型化的先进有效的方法。

7.2.2　数据挖掘的常用方法

利用数据挖掘进行数据分析常用的方法主要有分类、回归分析、聚类、关联规则、特征、变化和偏差分析、Web 页挖掘等，它们分别从不同的角度对数据进行挖掘。

1. 分类

分类是找出数据库中一组数据对象的共同特点并按照分类模式将其划分为不同的类，其目的是通过分类模型，将数据库中的数据项映射到某个给定的类别。它可以应用到客户的分类、客户的属性和特征分析、客户满意度分析、客户的购买趋势预测等，如一个汽车零售商将客户按照对汽车的喜好划分成不同的类，这样营销人员就可以将新型汽车的广告手册直接邮寄到有这种喜好的客户手中，从而大大增加了商业机会。

2. 回归分析

回归分析方法反映的是事务数据库中属性值在时间上的特征，产生一个将数据项映

射到一个实值预测变量的函数，发现变量或属性间的依赖关系，其主要研究问题包括数据序列的趋势特征、数据序列的预测以及数据间的相关关系等。它可以应用到市场营销的各个方面，如客户寻求、保持和预防客户流失活动、产品生命周期分析、销售趋势预测及有针对性的促销活动等。

3. 聚类

聚类分析是把一组数据按照相似性和差异性分为几个类别，其目的是使得属于同一类别的数据间的相似性尽可能大，不同类别中的数据间的相似性尽可能小。它可以应用到客户群体的分类、客户背景分析、客户购买趋势预测、市场的细分等。

4. 关联规则

关联规则是描述数据库中数据项之间所存在的关系的规则，即根据一个事务中某些项的出现可导出另一些项在同一事务中也出现，即隐藏在数据间的关联或相互关系。在客户关系管理中，通过对企业的客户数据库里的大量数据进行挖掘，可以从大量的记录中发现有趣的关联关系，找出影响市场营销效果的关键因素，为产品定位、定价与定制客户群，客户寻求、细分与保持，市场营销与推销，营销风险评估和诈骗预测等决策支持提供参考依据。

5. 特征

特征分析是从数据库中的一组数据中提取出关于这些数据的特征式，这些特征式表达了该数据集的总体特征。如营销人员通过对客户流失因素的特征提取，可以得到导致客户流失的一系列原因和主要特征，利用这些特征可以有效地预防客户的流失。

6. 变化和偏差分析

偏差包括很大一类潜在有趣的知识，如分类中的反常实例、模式的例外、观察结果对期望的偏差等，其目的是寻找观察结果与参照量之间有意义的差别。在企业危机管理及其预警中，管理者更感兴趣的是那些意外规则。意外规则的挖掘可以应用到各种异常信息的发现、分析、识别、评价和预警等方面。

7. Web 页挖掘

随着 Internet 的迅速发展及 Web 的全球普及，使得 Web 上的信息量无比丰富，通过对 Web 的挖掘，可以利用 Web 的海量数据进行分析，收集政治、经济、政策、科技、金融、各种市场、竞争对手、供求信息、客户等有关的信息，集中精力分析和处理那些对企业有重大或潜在重大影响的外部环境信息和内部经营信息，并根据分析结果找出企业管理过程中出现的各种问题和可能引起危机的先兆，对这些信息进行分析和处理，以便识别、分析、评价和管理危机。

7.2.3　数据挖掘的功能

数据挖掘通过预测未来趋势及行为，做出前摄的、基于知识的决策。数据挖掘的目标是从数据库中发现隐含的、有意义的知识，主要有以下五类功能。

1. 自动预测趋势和行为

数据挖掘自动在大型数据库中寻找预测性信息，以往需要进行大量手工分析的问题如今可以迅速直接由数据本身得出结论。一个典型的例子是市场预测问题，数据挖掘使用过去有关促销的数据来寻找未来投资中回报最大的用户，其他可预测的问题包括预报破产以及认定对指定事件最可能做出反应的群体。

2. 关联分析

数据关联是数据库中存在的一类重要的可被发现的知识。若两个或多个变量的取值之间存在某种规律性，就称为关联。关联可分为简单关联、时序关联、因果关联。关联分析的目的是找出数据库中隐藏的关联网。有时并不知道数据库中数据的关联函数，即使知道也是不确定的，因此关联分析生成的规则带有可信度。

3. 聚类

数据库中的记录可被划分为一系列有意义的子集，即聚类。聚类增强了人们对客观现实的认识，是概念描述和偏差分析的先决条件。聚类技术主要包括传统的模式识别方法和数学分类学。20 世纪 80 年代初，Mchalski 提出了概念聚类技术及其要点：在划分对象时不仅考虑对象之间的距离，还要求划分出的类具有某种内涵描述，从而避免了传统技术的某些片面性。

4. 概念描述

概念描述就是对某类对象的内涵进行描述，并概括这类对象的有关特征。概念描述分为特征性描述和区别性描述，前者描述某类对象的共同特征，后者描述不同类对象之间的区别。生成一个类的特征性描述只涉及该类对象中所有对象的共性。生成区别性描述的方法很多，如决策树方法、遗传算法等。

5. 偏差检测

数据库中的数据常有一些异常记录，从数据库中检测这些偏差很有意义。偏差包括很多潜在的知识，如分类中的反常实例、不满足规则的特例、观测结果与模型预测值的偏差、量值随时间的变化等。偏差检测的基本方法是，寻找观测结果与参照值之间有意义的差别。

7.2.4 数据挖掘技术

下面介绍数据挖掘的一些常用技术。

1. 人工神经网络

人工神经网络（Artificial Neural Network，ANN）是 20 世纪 80 年代以来人工智能领域兴起的研究热点。它从信息处理角度对人脑神经元网络进行抽象，建立某种简单模型，按不同的连接方式组成不同的网络。在工程与学术界也常直接简称为神经网络或类神经网络。神经网络是一种运算模型，由大量的结点（或称神经元）之间相互连接构成。每个结点代表一种特定的输出函数，称为激励函数（activation function）。每两个结点间的连

接都代表一个对于通过该连接信号的加权值,称为权重,这相当于人工神经网络的记忆。网络的输出则依网络的连接方式,权重值和激励函数的不同而不同。而网络自身通常都是对自然界某种算法或者函数的逼近,也可能是对一种逻辑策略的表达。

最近十多年来,人工神经网络的研究工作不断深入,已经取得了很大的进展,其在模式识别、智能机器人、自动控制、预测估计、生物、医学、经济等领域已成功地解决了许多现代计算机难以解决的实际问题,表现出了良好的智能特性。

2. 决策树

决策树(Decision Tree)是在已知各种情况发生概率的基础上,通过构成决策树来求取净现值的期望值大于等于零的概率,评价项目风险,判断其可行性的决策分析方法,是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干,故称决策树。在机器学习中,决策树是一个预测模型,他代表的是对象属性与对象值之间的一种映射关系。Entropy＝系统的凌乱程度,使用算法 ID3、C4.5 和 C5.0 生成树算法使用熵。这一度量是基于信息学理论中熵的概念。

决策树是一种树形结构,其中每个内部结点表示一个属性上的测试,每个分支代表一个测试输出,每个叶结点代表一种类别。

分类树(决策树)是一种十分常用的分类方法。它是一种监管学习,所谓监管学习,就是给定一堆样本,每个样本都有一组属性和一个类别,这些类别是事先确定的,那么通过学习得到一个分类器,这个分类器能够对新出现的对象给出正确的分类。这样的机器学习就被称为监管学习。

3. 遗传算法

遗传算法(Genetic Algorithm)是模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模型,是一种通过模拟自然进化过程搜索最优解的方法。

遗传算法是从代表问题可能潜在的解集的一个种群(population)开始的,而一个种群则由经过基因(gene)编码的一定数目的个体(individual)组成。每个个体实际上是染色体(chromosome)带有特征的实体。染色体作为遗传物质的主要载体,即多个基因的集合,其内部表现(即基因型)是某种基因组合,它决定了个体的形状的外部表现,如黑头发的特征是由染色体中控制这一特征的某种基因组合决定的。因此,在一开始需要实现从表现型到基因型的映射即编码工作。

由于仿照基因编码的工作很复杂,所以往往对其进行简化,如二进制编码,初代种群产生之后,按照适者生存和优胜劣汰的原理,逐代(generation)演化产生出越来越好的近似解。在每一代,根据问题域中个体的适应度(fitness)大小选择(selection)个体,并借助于自然遗传学的遗传算子(genetic operators)进行组合交叉(crossover)和变异(mutation),产生出代表新的解集的种群。这个过程将导致种群像自然进化一样的后生代种群比前代更加适应于环境,末代种群中的最优个体经过解码(decoding),可以作为问题近似最优解。

4. 邻近算法

邻近算法,或者说 k 最近邻(kNN,k-Nearest Neighbor)分类算法是数据挖掘分类技

术中最简单的方法之一。所谓 k 最近邻，就是 k 个最近的邻居的意思，说的是每个样本都可以用它最接近的 k 个邻居来代表。邻近算法如图 7.7 所示。

kNN 算法的核心思想是如果一个样本在特征空间中的 k 个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。该方法在确定分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

kNN 方法在类别决策时，只与极少量的相邻样本有关。由于 kNN 方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，kNN 方法较其他方法更为适合。

如图 7.8 所示，圆要被决定赋予哪个类，是三角形还是四方形？如果 $k=3$，由于三角形所占比例为 2/3，圆将被赋予三角形那个类，如果 $k=5$，由于四方形比例为 3/5，因此圆被赋予四方形类。

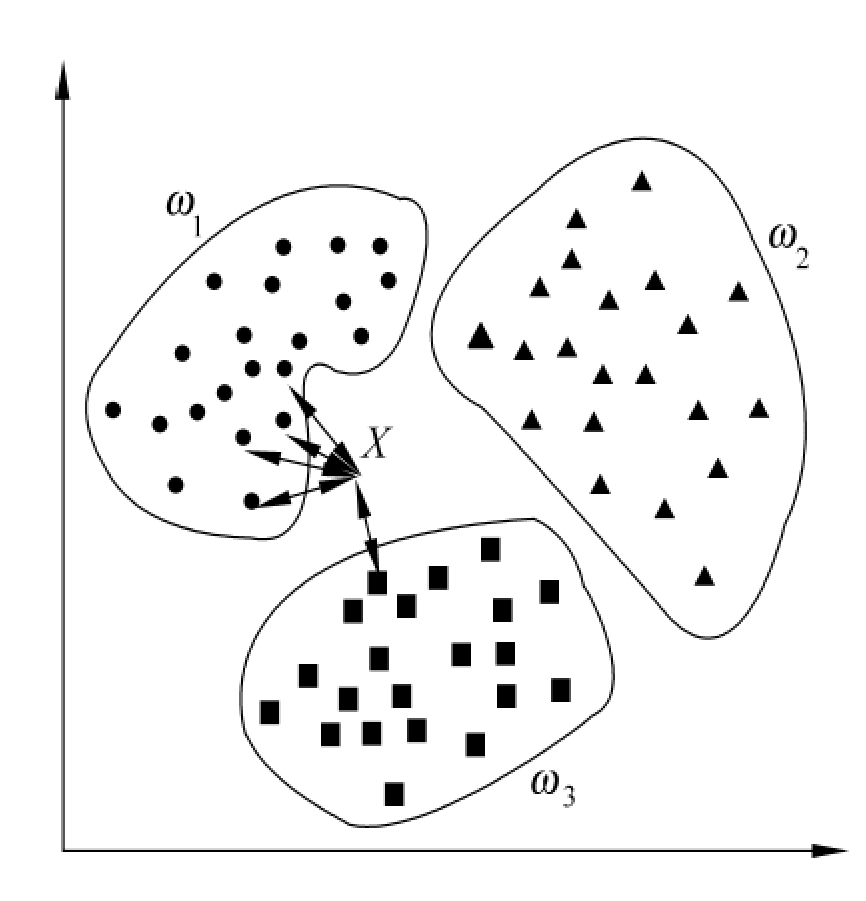


图 7.7 邻近算法

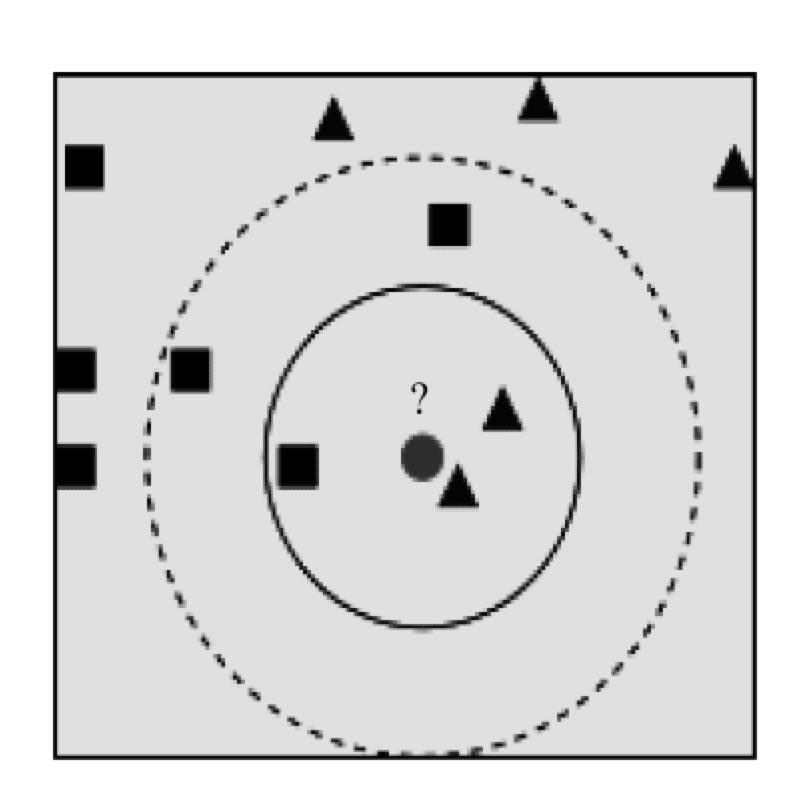


图 7.8 kNN 算法的决策过程

k 最近邻（k-Nearest Neighbor，kNN）分类算法，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。该方法的思路是：如果一个样本在特征空间中的 k 个最相似（即特征空间中最邻近）的样本中的大多数属于某一个类别，则该样本也属于这个类别。kNN 算法中，所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。kNN 方法虽然从原理上也依赖于极限定理，但在类别决策时，只与极少量的相邻样本有关。由于 kNN 方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，kNN 方法较其他方法更为适合。

kNN 算法不仅可以用于分类，还可以用于回归。通过找出一个样本的 k 个最近邻居，将这些邻居的属性的平均值赋给该样本，就可以得到该样本的属性。更有用的方法是将不同距离的邻居对该样本产生的影响给予不同的权值（weight），如权值与距离成反比。

7.2.5 数据挖掘的流程

1. 数据挖掘环境

数据挖掘是指一个完整的过程，该过程从大型数据库中挖掘先前未知的、有效的、可

实用的信息,并使用这些信息做出决策或丰富知识。

2. 数据挖掘过程图

图 7.9 描述了数据挖掘的基本过程和主要步骤。

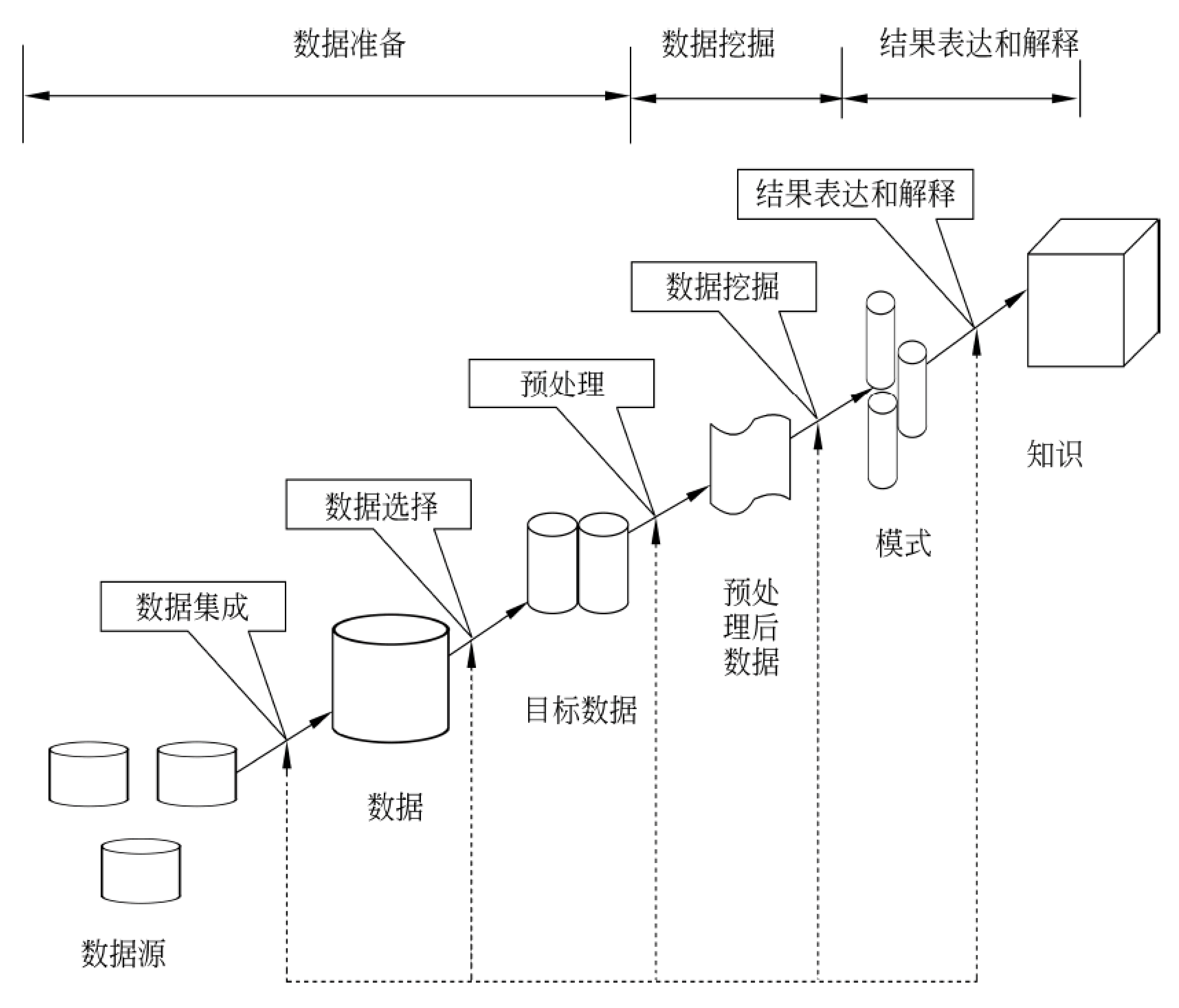


图 7.9　典型数据挖掘系统的过程

3. 数据挖掘过程工作量

在数据挖掘中被研究的业务对象是整个过程的基础,它驱动了整个数据挖掘过程,也是检验最后结果和指引分析人员完成数据挖掘的依据和顾问。各个步骤是按一定顺序完成的,当然整个过程中还会存在步骤间的反馈。数据挖掘的过程并不是自动的,绝大多数的工作需要人工完成。各步骤在整个过程中的工作量之比。可以看到,60%的时间用在数据准备上,这说明了数据挖掘对数据的严格要求,而后挖掘工作仅占总工作量的 10%。

4. 数据挖掘过程简介

过程中各个步骤的大体内容如下:

1) 确定业务对象

清晰地定义出业务问题,认清数据挖掘的目的是数据挖掘的重要一步。挖掘的最后结构是不可预测的,但要探索的问题应是有预见的,为了数据挖掘而数据挖掘则带有盲目性,是不会成功的。

2) 数据准备

(1) 数据的选择。

搜索所有与业务对象有关的内部和外部数据信息,并从中选择出适用于数据挖掘应

用的数据。

(2) 数据的预处理。

研究数据的质量,为进一步的分析做准备,并确定将要进行的挖掘操作的类型。

(3) 数据的转换。

将数据转换成一个分析模型。这个分析模型是针对挖掘算法建立的。建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键。

3) 数据挖掘

对所得到的经过转换的数据进行挖掘。除了完善从选择合适的挖掘算法外,其余一切工作都能自动完成。

4) 结果分析

解释并评估结果。其使用的分析方法一般应作数据挖掘操作而定,通常会用到可视化技术。

5) 知识的同化

将分析所得到的知识集成到业务信息系统的组织结构中去。

5. 数据挖掘需要的人员

数据挖掘过程的分步实现,不同的步骤需要有不同专长的人员,大体可以分为三类。

业务分析人员:要求精通业务,能够解释业务对象,并根据各业务对象确定出用于数据定义和挖掘算法的业务需求。

数据分析人员:精通数据分析技术,并对统计学有较熟练的掌握,有能力把业务需求转化为数据挖掘的各步操作,并为每步操作选择合适的技术。

数据管理人员:精通数据管理技术,并从数据库或数据仓库中收集数据。

从上可见,数据挖掘是一个多种专家合作的过程,也是一个在资金上和技术上高投入的过程。这一过程要反复进行并在反复过程中,不断地趋近事物的本质,不断地优化问题的解决方案。

7.2.6 数据挖掘的应用

1. 数据挖掘解决的典型商业问题

需要强调的是,数据挖掘技术从一开始就是面向应用的。目前,在很多领域,数据挖掘(data mining)都是一个很时髦的词,尤其是在如银行、电信、保险、交通、零售(如超级市场)等商业领域。数据挖掘所能解决的典型商业问题包括数据库营销(Database Marketing)、客户群体划分(Customer Segmentation & Classification)、背景分析(Profile Analysis)、交叉销售(Cross-selling)等市场分析行为,以及客户流失性分析(churn Analysis)、客户信用记分(Credit Scoring)、欺诈发现(Fraud Detection)、故障诊断等等。

2. 数据挖掘在市场营销的应用

数据挖掘技术在企业市场营销中得到了比较普遍的应用,它以市场营销学的市场细分原理为基础,其基本假定是"消费者过去的行为是其今后消费倾向的最好说明"。

通过收集、加工和处理涉及消费者消费行为的大量信息,确定特定消费群体或个体的

兴趣、消费习惯、消费倾向和消费需求，进而推断出相应消费群体或个体下一步的消费行为，然后以此为基础，对所识别出来的消费群体进行特定内容的定向营销，这与传统的不区分消费者对象特征的大规模营销手段相比，大大节省了营销成本，提高了营销效果，从而为企业带来更多的利润。

3. 案例——信用卡消费的数据挖掘

商业消费信息来自市场中的各种渠道。例如，每当用信用卡消费时，商业企业就可以在信用卡结算过程收集商业消费信息，记录下人们进行消费的时间、地点、感兴趣的商品或服务、愿意接收的价格水平和支付能力等数据；当我们在申办信用卡、办理汽车驾驶执照、填写商品保修单等其他需要填写表格的场合时，我们的个人信息就存入了相应的业务数据库；企业除了自行收集相关业务信息之外，甚至可以从其他公司或机构购买此类信息为自己所用。

这些来自各种渠道的数据信息被组合，应用超级计算机、并行处理、神经元网络、模型化算法和其他信息处理技术手段进行处理，从中得到商家用于向特定消费群体或个体进行定向营销的决策信息。这种数据信息是如何应用的呢？

举一个简单的例子。当银行通过对业务数据进行挖掘后，发现一个银行账户持有者突然要求申请双人联合账户时，并且确认该消费者是第一次申请联合账户，银行会推断该用户可能要结婚了，它就会向该用户定向推销用于购买房屋、支付子女学费等长期投资业务，银行甚至可能将该信息卖给专营婚庆商品和服务的公司。数据挖掘构筑竞争优势。

在市场经济比较发达的国家和地区，许多公司都开始在原有信息系统的基础上通过数据挖掘对业务信息进行深加工，以构筑自己的竞争优势，扩大自己的营业额。

美国运通公司(American Express)有一个用于记录信用卡业务的数据库，数据量达到 54 亿字符，并仍在随着业务进展不断更新。运通公司通过对这些数据进行挖掘，制定了“关联结算(Relation ship Billing)优惠”的促销策略，即如果一个顾客在一个商店用运通卡购买一套时装，那么在同一个商店再买一双鞋，就可以得到比较大的折扣，这样既可以增加商店的销售量，也可以增加运通卡在该商店的使用率。

7.2.7　“大数据自动挖掘”才是大数据的真正意义

1. 大数据不是指很多数据

“大数据”只是个简称，说全一点应是“大数据挖掘”，没经过挖掘的大数据只是没有开采出来的原油，一点用处都没有。

2. 大数据也不是指一般意义上的数据挖掘

有很多人以前是搞数据分析或数据挖掘的，当《大数据时代》这本书一问世、大数据开始火的时候，他们摇身一变就成了搞大数据的专家了。如果真是这样，就根本没必要提大数据这事儿，因为它本来就一直存在着，只不过换个说法。就好像我们没必要今天突然提出个“饮 H_2O”的说法来代替“喝水”。嗯，对，那叫玩概念。

3. “大数据挖掘”其实还没有说全，再说完整点，应该是“大数据自动挖掘”

以前的数据分析或挖掘，是指人通过数据去进行分析，挖掘出一些规律性的东西以供

以后使用。

但面对大数据，由于不光是数据量太大，而且往往包括数据的维度也很多，人已不可能去处理这样海量的数据，甚至如何处理都不知道，这时必须用计算机来自动处理，挖掘出数据中的规律。

但是目前计算机还不能像人那样进行严密、复杂的逻辑思维，因此它们也无法用人的思维模式去分析数据，人可能只要较少的数据就能分析出其中的规律，数据多了反而没有办法，所以我们人类都是采用抽样分析。

计算机则正好相反，无法根据少量数据去分析出规律，但它有一个优势，那就是运算速度非常快，因此有可能处理海量数据以后找出其中的规律。

由于计算机还不能进行复杂的逻辑思维，所以它的处理方法很简单，就是进行简单的统计运算，也就是“硬算”，统计出在什么情况会出什么样的结果，然后当类似的情况再出现时，它就会告诉我们可能会出现某种结果了。

由这里也可看大数据的另一个特点，即大数据主要是进行预测，告诉你未来将会出现什么样的结果。而不是只分析出过去的走势和现状，未来还是要由人去判断。

为什么这种简单的方法会有效呢？这就回到“大数据”这个词上来了，那就是因为数据量非常大，统计出来的结果就往往是正确的。

大家一定都知道这个例子，扔硬币来统计正、反面出现的几率，如果只扔 10 次，也许正面出现 9 次，以此来得出结论肯定是错的；但如果你扔 10 万次、100 万次，甚至更多，那你统计出来的结果基本是正确的，正、反面出现的几率一定是各 50%。

是的，大数据自动挖掘就是依据这一原理。

这里没有严密的因果分析，不是通过数据分析出原因再推导出结果；而是通过统计知道有这样的情况，一般就会有这样的结果，也即现象与结果的相关性。所以大数据就有一个显著的特点，只关心相关性，不关心因果；用更通俗的话说就是“只知道结果，不知道原因”。

这实际是人们根据电脑的优势，找出了一个全新的数据分析、挖掘方式，与传统的方式完全不同。

7.3 商业智能与数据分析

7.3.1 商业智能技术辅助决策的发展

商务智能，英文为 Business Intelligence，简写为 BI。

商业智能的概念在 1996 年最早由加特纳集团(Gartner Group)提出，加特纳集团将商业智能定义为：商业智能描述了一系列的概念和方法，通过应用基于事实的支持系统来辅助商业决策的制定。商业智能技术提供使企业迅速分析数据的技术和方法，包括收集、管理和分析数据，将这些数据转化为有用的信息，然后分发到企业各处。

商业智能通常被理解为将企业中现有的数据转化为知识，帮助企业做出明智的业务经营决策的工具。这里所谈的数据包括来自企业业务系统的订单、库存、交易账目、客户

和供应商等来自企业所处行业和竞争对手的数据以及来自企业所处的其他外部环境中的各种数据。而商业智能能够辅助进行的业务经营决策,既可以是操作层的,也可以是战术层和战略层的决策。为了将数据转化为知识,需要利用数据仓库、联机分析处理(OLAP)工具和数据挖掘等技术。因此,从技术层面上讲,商业智能不是什么新技术,它只是数据仓库、OLAP 和数据挖掘等技术的综合运用。

可以认为,商业智能是对商业信息的搜集、管理和分析过程,目的是使企业的各级决策者获得知识或洞察力(insight),促使他们做出对企业更有利的决策。商业智能一般由数据仓库、联机分析处理、数据挖掘、数据备份和恢复等部分组成。商业智能的实现涉及到软件、硬件、咨询服务及应用,其基本体系结构包括数据仓库、联机分析处理和数据挖掘三个部分。

因此,把商业智能看成是一种解决方案应该比较恰当。商业智能的关键是从许多来自不同的企业运作系统的数据中提取出有用的数据并进行清理,以保证数据的正确性,然后经过抽取(Extraction)、转换(Transformation)和装载(Load),即 ETL 过程,合并到一个企业级的数据仓库里,从而得到企业数据的一个全局视图,在此基础上利用合适的查询和分析工具、数据挖掘工具(大数据魔镜)、OLAP 工具等对其进行分析和处理(这时信息变为辅助决策的知识),最后将知识呈现给管理者,为管理者的决策过程提供支持。

提供商业智能解决方案的著名 IT 厂商包括微软、IBM、Oracle、SAP、Informatica、Microstrategy、SAS、Royalsoft 等。

7.3.2　商业智能系统架构

从系统的观点来看,商业智能的过程是这样的:从不同的数据源收集的数据中提取有用的数据,对数据进行清理以保证数据的正确性,将数据经转换、重构后存入数据仓库或数据场(这时数据变为信息),然后寻找合适的查询和分析工具,数据挖掘工具,OLAP 工具对信息进行处理(这时信息变为辅助决策的知识),最后将知识呈现于用户面前,转变为决策。可以看出,商业智能最大限度地利用了企业操作系统(ERP)中的数据,将数据整理为信息,再升华为知识,所以对用户提供了最大程度的支持。

7.3.3　商业智能的技术体系

商业智能的技术体系主要由数据仓库(DW)、在线分析处理(OLAP)以及数据挖掘(DM)三部分组成。商业智能中所包含的数据分析技术主要可分为以下三个阶段。

1. 数据仓库(Data Warehouse)

为了有效地进行营销管理,企业往往需要将各地的数据汇总到总部,并建立一个庞大的数据仓库。这种数据仓库不但能够保存历史数据、阶段性数据,并从时间上进行分析,而且能够装载外部数据,接受大量的外部查询。

建立数据仓库的过程一般包括清洗、抽取数据操作,统一数据格式,设定自动程序以定时抽取操作数据并自动更新数据仓库,预先执行合计计算等步骤。

快速、简单、易用的查询和报告工具能够帮助管理者充分利用企业中不同层次的数

据，获取所需要的特定信息，并以合理的格式加以显示。同时，优秀的工具支持多种网络环境，允许用户在客户机/服务器网络、内部网络或 Internet 上传输分析结果。它们还应该有足够的灵活性，以支持各种类型的查询和报告需求，从简单的订阅、周期性的报告，到使用 SQL 和其他查询语言作随机查询。

2. 在线分析处理（OLAP）

在线分析处理是一种高度交互式的过程，信息分析专家可以即时进行反复分析，迅速获得所需结果。在线分析处理同时也是对存储在多维数据库（MDD）或关系型数据库（RDBMS）中的数据进行分析、处理的过程。这种分析可以是多维在线分析处理、关系型在线分析处理，也可以是混合在线分析处理。

这一过程一般包括三种可供选择的方案：

- 预先计算——小结数据在使用前进行计算并存储；
- 即时计算和存储——小结数据在查询是计算，然后存储结果。因为消除了相应的运行计算，使随后的查询运行变得更快。
- 随时计算——用户在需要时对小结数据进行计算。

3. 数据挖掘（Data Mining）

数据挖掘是从浩如瀚海的数据和文档中发现以前未知的、可以理解的信息的过程。由于数据挖掘的价值在于扫描数据仓库或建立非常复杂的查询，数据和文本挖掘工具必须提供很高的吞吐量，并拥有并行处理功能，而且可以支持多种采集技术。数据挖掘工具应该拥有良好的扩展功能，并且能够支持将来可能遇到的各种数据（或文档）和计算环境。

4. 总结

商业智能是帮助客户将数据转化为利润的手段。实质上，商业智能就是帮助企业充分利用已有数据，将其分析整理为可用信息，并以此作为企业决策的依据。

目前，多数企业在部署系统时多针对自身当前的业务需求，着眼于静态的处理，无法有效地预测即将产生的情况。在这种条件下，他们难免处于被动的边缘，在市场的波澜面前仓促做出应对之策，其效果自然就可想而知了。企业若想改变一直面临的被动局面，就必须利用智能的解决方案，高效地收集、整理并分析相关数据，为企业的正确决策提供前瞻性支持。

7.3.4 商务智能＝数据＋分析＋决策＋利益

1. 背景介绍

人类社会从物物交换到货币的产生，到形形色色的交易，产生了现在繁荣、复杂的各种商业活动。利益是商务的核心，而商务需要经过买卖双方的交易，谈判，而商品的流通又需要物流、库存，其中业务流程十分烦琐，然而科技进步改善或者正在改变着其形式，人们的工作效率正在极大地提高。

在这个信息化的时代，许多传统业务被信息化手段所取代或者信息化作为其辅助手段。于是，在这个时代，所有的人都在谈数据，并且相关的商务数据呈爆炸性指数级的增

长。可是，不是所有的数据都是有用的，所以人们需要从中挖掘有用的信息，用于指导现实工作。

商务智能通常被理解为将企业中现有的数据转化为知识，帮助企业做出明智的业务经营决策的工具。比如，百货商场每天有各种各样的商品被出售，其 POS 系统存储着商品的销售情况，数据量十分庞大。在这些数据基础上，利用一定的数学模型和智能软件工具进行分析，知道哪些产品最热销，哪些时段人们喜欢购买什么。

接着，运用分析后的结果进行决策，比如分析后得知下雨天的时候啤酒和炸鸡的销量比其他天气时段更多，于是我们决定在下雨的日子增大啤酒和炸鸡的产量。通过这些分析和决策，得到了商业利润的增加，这种利润是利用现代工具进行商务智能活动的动力。这个过程可以总结为以下的一个等式：

商务智能＝数据＋分析＋决策＋利益

2. 数据获取

传统的数据获取是手工进行纸质记录，缺点是记录容易出错，且随着时间的流动，其数量会大大增加以致于查找历史数据的困难。比如，传统地主家的管家进行家庭财政的登记，账本厚又重，对账极其麻烦，而且说不定账本会因为火灾或各种原因而破损，如被老鼠咬烂了。

随着科技的进步，有了计算机，于是数据存到了磁带，然后是磁盘。世界因有了社会分工而变得美妙，每个人都在自己擅长的领域工作，从而创造着更大的利益。于是，不懂计算机的人借助着别人开发的管理系统进行数据的管理，比如超市的商品管理系统、公司内部的人员管理系统。而软件程序员借助了数据库、数据仓库等产品进行设计编码，创造了上述的管理系统。

于是，一层接力一层，数据的获取从手工一个个用笔记下来到使用计算机键盘进行录入。通过现代科技手段，查看历史数据只要进行搜索，很快很好就能得到十年前的数据，从而可以更高效率地进行数据分析。

商务智能，智能二字凸显了计算机的重要性。计算机的一切都是由 0、1 二进制数组成，这两个最普通不过的符号构建了计算机整个数据大厦。如何更好地将数据存到计算机磁盘中，并迅速读取出来呢？早期的数据存储是使用卡片进行数据读取，后来便产生了现代计算机的存储体系、寄存器、内存、磁盘。从硬件开始，后来出现了软件层面的文件系统、IO 流。为了更便于存储大量数据，出现了数据库软件，各种数据库理论和工具开始出现。

目前使用最多的数据库是 1993 年 E. F. Codd 提出的关系数据库。

3. 数据分析

数据分析方面主要依赖数据挖掘方面的知识，因为商务智能是数据挖掘领域的一个分支。数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。

数据挖掘利用了来自如下一些领域的思想：

(1) 来自统计学的抽样、估计和假设检验。

(2) 人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。

数据挖掘也迅速地接纳了来自其他领域的思想,这些领域包括最优化、进化计算、信息论、信号处理、可视化和信息检索。一些其他领域也起到重要的支撑作用。特别地,需要数据库系统提供有效的存储、索引和查询处理支持。源于高性能(并行)计算的技术在处理海量数据集方面常常是重要的。分布式技术也能帮助处理海量数据,并且当数据不能集中到一起处理时更是至关重要。

主要的分析算法有分类(Classification)估计(Estimation)预测(Prediction)相关性分组或关联规则(Affinity grouping or association rules)聚类(Clustering)等。这些算法主要依赖数学进行构建,大多数商业数据挖掘软件已经实现了这些功能,方便普通人士的使用。

通过使用数据挖掘软件,可以对存储在数据库中的数据进行分析处理,得到一定的统计和计算结果。这些结果可以指导现实的决策。

目前的数据挖掘软件有用于一般分析目的的软件包 SAS Enterprise Miner、SPSS Clementine 和 IBM Intelligent Miner 等,还有针对特定功能或产业而研发的软件,如 KD1(针对零售业)、Options & Choices(针对保险业)、HNC(针对信用卡诈欺或呆账侦测)、Unica Model 1(针对行销业)、iEM System(针对流程行业的实时历史数据)等。

4. 商务决策

随着数据库技术的发展和应用,数据库存储的数据量从 20 世纪 80 年代的兆(M)字节及千兆(G)字节过渡到现在的太(T)字节和拍(P)字节,同时,用户的查询需求也越来越复杂,涉及的已不仅是查询或操纵一张关系表中的一条或几条记录,而且要对多张表中千万条记录的数据进行数据分析和信息综合,关系数据库系统已不能全部满足这一要求。在国外,不少软件厂商采取了发展其前端产品来弥补关系数据库管理系统支持的不足,力图统一分散的公共应用逻辑,在短时间内响应非数据处理专业人员的复杂查询要求。

联机分析处理(OLAP)系统是数据仓库系统最主要的应用,专门设计用于支持复杂的分析操作,侧重对决策人员和高层管理人员的决策支持,可以根据分析人员的要求快速、灵活地进行大数据量的复杂查询处理,并且以一种直观而易懂的形式将查询结果提供给决策人员,以便他们准确掌握企业(公司)的经营状况,了解对象的需求,制定正确的方案。

OLAP 工具是针对特定问题的联机数据访问与分析。它通过多维的方式对数据进行分析、查询和生成报表。维是人们观察数据的特定角度。例如,一个企业在考虑产品的销售情况时,通常从时间、地区和产品的不同角度来深入观察产品的销售情况。

这里的时间、地区和产品就是维。而这些维的不同组合和所考察的度量指标构成的多维数组则是 OLAP 分析的基础,可形式化表示为(维 1,维 2,…,维 n,度量指标),如(地区、时间、产品、销售额)。多维分析是指对以多维形式组织起来的数据采取切片(Slice)、切块(Dice)、钻取(Drill-down 和 Roll-up)、旋转(Pivot)等各种分析动作,以求剖析数据,使用户能从多个角度、多侧面地观察数据库中的数据,从而深入理解包含在数据中的

信息。

商务决策使用了上述的数据挖掘软件得出的结果，而 OLAP 是一个更加方便的系统，能更快、更好地将分析的结果以图表等方式进行展示，方便决策人员进行对比、讨论。通过智能化工具的处理，领导和改革者可以决定是否开展某项业务，或者如何进行某项业务，这也是称之为商务决策的原因。

5. 利益动力

商业智能的关键是从许多来自不同的企业运作系统的数据中提取出有用的数据并进行清理，以保证数据的正确性，然后经过抽取（Extraction）、转换（Transformation）和装载（Load），即 ETL 过程，合并到一个企业级的数据仓库里，从而得到企业数据的一个全局视图，在此基础上利用合适的查询和分析工具、数据挖掘工具、OLAP 工具等对其进行分析和处理（这时信息变为辅助决策的知识），最后将知识呈现给管理者，为管理者的决策过程提供支持。

商务智能＝数据＋分析＋决策＋利益，等式包含了利益，是因为利益作为一种动力，促进了商务智能的发展。因为想改变，所以改变。因为想提高效率，所以改变。因为要以最小的投入挣得最大的利益，所以要改变。人类生活的改变来源人类对美好生活的追求，想把人类从繁忙的体力劳动中解放出来。计算机这一科技产物，与商务联系起来，必定能够创造极大的价值。

7.4　电商大数据分析技术

7.4.1　移动互联网应用数据分析基础

现在诸多大型互联网公司其移动端的流量已经超越 PC 端的流量，很多大型互联网企业 PC 业务用户往移动端迁移，呈现出 PC 业务增长放缓、移动业务增长迅速的态势。从第三方数据机构统计的数据来看，网民中使用手机上网的人群占比进一步提升，由 2013 年 12 月的 81.0%提升至 2015 年 6 月的 88.9%，即中国网民中，接近 9 成的用户在使用手机上网，达到接近 6 亿的规模。如果一个互联网企业没有在移动端的拳头产品，将很快被移动互联网的浪潮颠覆。

中国互联网网民规模的统计如图 7.10 所示。

从数据看出，移动互联网是互联网发展最重要的方向，因此，对于拥抱互联网的企业来说，设计和运营好移动互联网应用（以下称 APP）成为移动互联网时代最重要的任务。而在移动互联网的设计和运营过程中，数据分析起到很基础但也很重要的作用。在互联网企业，任何一个 APP 都要事先规划好数据体系，才允许上线运营，有了数据才可以更好地科学运营。下面将为大家介绍 APP 的基础数据指标体系。

APP 的数据指标体系主要分为五个维度，包括用户规模与质量、参与度分析、渠道分析、功能分析以用户属性分析。用户规模和质量维度主要是分析用户规模指标，这类指标一般为产品考核的重点指标；参与度分析主要分析用户的活跃度；渠道分析主要分析渠道推广效果；功能分析主要分析功能活跃情况、页面访问路径以及转化率；用户属性分析主

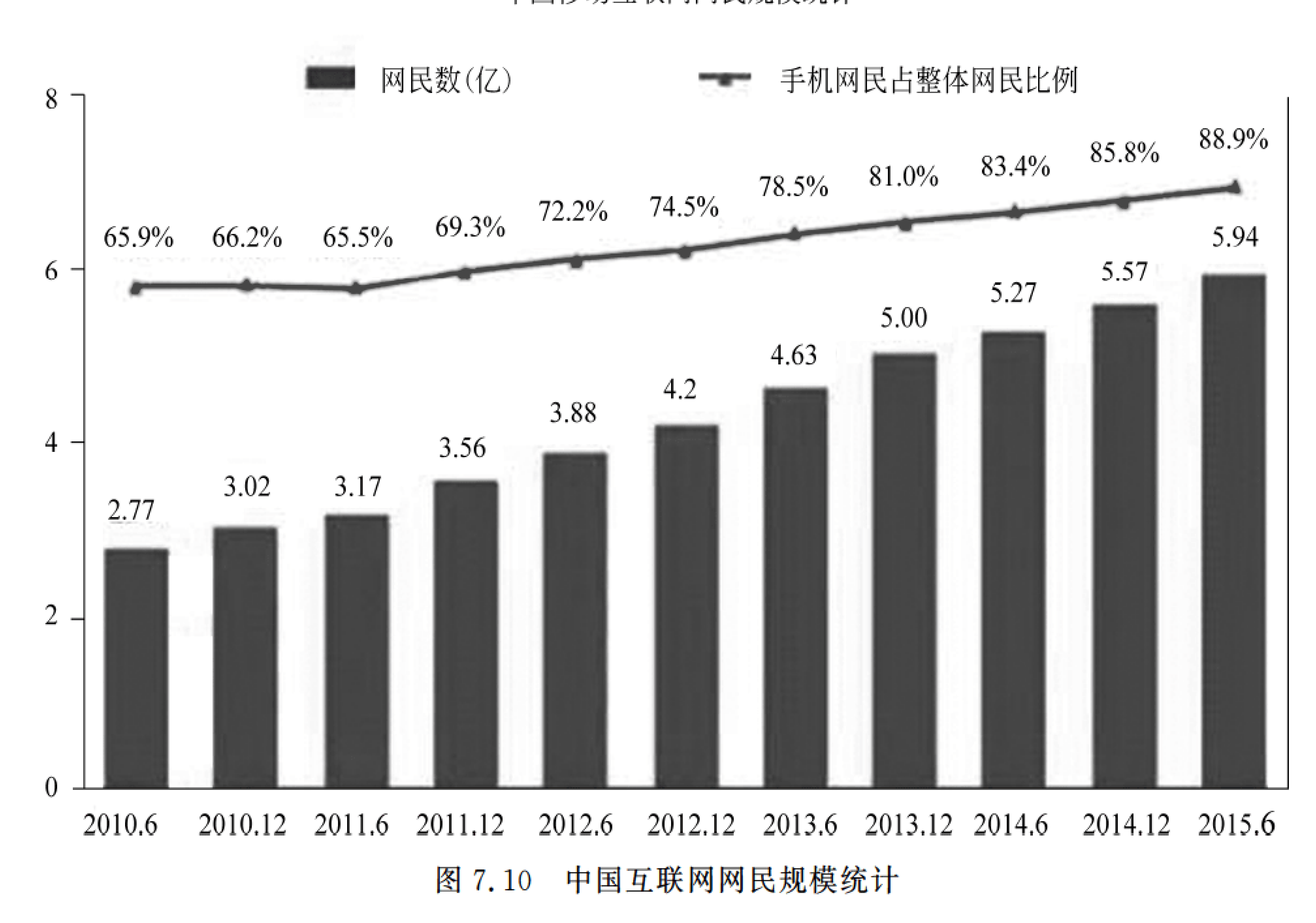


图 7.10　中国互联网网民规模统计

要分析用户特征。

7.4.2　用户规模和质量

用户规模和质量的分析包括活跃用户、新增用户、用户构成、用户留存率、每个用户总活跃天数五个常见指标。用户规模和质量是 APP 分析最重要的维度，其指标也是相对其他维度最多，产品负责人要重点关注这个维度的指标。

1. 活跃用户指标

活跃用户指在某统计周期内启动过应用(APP)的用户。活跃用户数一般按照设备维度统计，即统计一段周期内启动过的设备(如手机、平板电脑)数量。活跃用户是衡量应用用户规模的指标。通常，一个产品是否成功，如果只看一个指标，那么这个指标一定是活跃用户数。很多互联网企业对产品负责人的 KPI 考核指标都以活跃用户数作为考核指标。活跃用户数根据不同统计周期可以分为日活跃数(DAU)、周活跃数(WAU)、月活跃数(MAU)。

大多数希望用户每天都打开的应用如新闻 APP、社交 APP、音乐 APP 等，其产品的 KPI 考核指标均为日活跃用户数(DAU)。为什么？如果这些 APP 考核的指标是月活跃用户数，那么会出现什么状况？

月活跃用户只要求用户在一个月内启动应用一次既可以计算为月活跃用户，所以，一个本应该每天都要启动的应用，如果用月活跃用户数作为 KPI 来考核，那么会出现产品运营负责人“偷懒”的情况，产品运营人员只需要每月想办法让用户启动一次即可，也许向用户推送两三个活动就可以实现，这样的考核会导致产品不够吸引力甚至是不健康的。

如果用日活跃用户来作为 KPI 来考核这个产品，那么产品运营负责人一定会设计让用户每天都想用的功能或者更新每天用户都想看的内容来吸引用户来使用。

2. 新增用户指标

新增用户是指安装应用后，首次启动应用的用户。按照统计时间跨度不同分为日、周、月新增用户。新增用户量指标主要是衡量营销推广渠道效果的最基础指标；另一方面，新增用户占活跃用户的比例也可以用于衡量产品健康度。如果某产品新用户占比过高，那说明该产品的活跃是靠推广得来，这种情况非常值得关注，尤其是关注用户的留存率情况。

3. 用户构成指标

用户构成是对周活跃用户或者月活跃用户的构成进行分析，有助于通过新老用户结构了解活跃用户健康度。以周活跃用户为例，周活跃用户包括以下几类用户，包括本周回流用户、连续活跃 n 周用户、忠诚用户、连续活跃用户。

本周回流用户是指上周未启动过应用，本周启动应用的用户；连续活跃 n 周用户是指连续 n 周，每周至少启动过一次应用的活跃用户；忠诚用户是指连续活跃 5 周及以上的用户；连续活跃用户是指连续活跃 2 周及以上的用户；近期流失用户是指连续 n 周（大于等于 1 周，但小于等于 4 周）没有启动过应用但用户。

4. 用户留存率指标

用户留存率是指在某一统计时段内的新增用户数中再经过一段时间后仍启动该应用的用户比例。用户留存率可重点关注次日、7 日、14 日以及 30 日留存率。

次日留存率即某一统计时段（如今天）新增用户在第二天（如明天）再次启动应用的比例。

7 日留存率即某一统计时段（如今天）新增用户数在第 7 天再次启动该应用的比例。

14 日和 30 日留存率以此类推。

用户留存率是验证产品用户吸引力很重要的指标。通常，我们可以利用用户留存率对比同一类别应用中不同应用的用户吸引力。如果对于某一个应用，在相对成熟的版本情况下，如果用户留存率有明显变化，则说明用户质量有明显变化，很可能是因为推广渠道质量的变化所引起的。

5. 每个用户总活跃天数指标

每个用户的总活跃天数指标（Total Active Days per user，TAD）是在统计周期内，平均每个用户在应用的活跃天数。如果统计周期比较长，如统计周期一年以上，那么，每个用户的总活跃天数基本可以反映用户在流失之前在 APP 上耗费的天数，这是反映用户质量尤其是用户活跃度很重要的指标。

7.4.3　参与度分析

参与度分析的常见分析包括启动次数分析、使用时长分析、访问页面分析和使用时间间隔分析。参与度分析主要是分析用户的活跃度。

1. 启动次数指标

启动次数是指在某一统计周期内用户启动应用的次数。在进行数据分析时，一方面要关注启动次数的总量走势，另一方面，则需要关注人均启动次数，即同一统计周期的启动次数与活跃用户数的比值，如人均日启动次数，则为日启动次数与日活跃用户数的比值，反映的是每天每用户平均启动次数。通常，人均启动次数和人均使用时长可以结合在一起分析。

2. 使用时长

使用总时长是指在某一统计统计周期内所有从 APP 启动到结束使用的总计时长。使用时长还可以从人均使用时长、单次使用时长等角度进行分析。

人均使用时长是同一统计周期内的使用总时长和活跃用户数的比值；单次使用时长是同一统计周期内使用总时长和启动次数的比值。

使用时长相关的指标也是衡量产品活跃度、产品质量的重要指标，道理很简单，用户每天的时间是有限的且宝贵的，如果用户愿意在你的产品投入更多的时间，证明你的应用对用户很重要。启动次数和使用时长可以结合在一起分析，如果用户启动次数多，使用时长长，则该 APP 则为用户质量非常高，用户黏性好的应用，比如现在很流行的社交应用。

3. 访问页面

访问页面数指用户一次启动访问的页面数。我们通常要分析访问页面数分布，即统计一定周期内（如 1 天、7 天或 30 天）应用的访问页面数的活跃用户数分布，如访问 1～2 页的活跃用户数、3～5 页的活跃用户数、6～9 页的活跃用户数、10～29 页的活跃用户数、30～50 页的活跃用户数，以及 50 页以上的活跃用户数。同时，我们可以通过不同统计周期（但统计跨度相同，如都为 7 天）的访问页面分布的差异，以便于发现用户体验的问题。

4. 使用时间间隔

使用时间间隔是指同一用户相邻两次启动的时间间隔。我们通常要分析使用时间间隔分布，一般统计一个月内应用的用户使用时间间隔的活跃用户数分布，如使用时间间隔在 1 一天内、1 天、2 天……7 天、8～14 天、15～30 天的活跃用户数分布。同时，我们可以通过不同统计周期（但统计跨度相同，如都为 30 天）的使用时间间隔分布的差异，以便于发现用户体验的问题。

7.4.4 渠道分析

渠道分析主要是分析各渠道在相关的渠道质量的变化和趋势，以科学评估渠道质量，优化渠道推广策略。渠道分析需要渠道推广负责人重点关注，尤其是目前移动应用市场渠道作弊较为盛行的情况下，渠道推广的分析尤其是要重点关注渠道作弊的分析。

渠道分析包括新增用户、活跃用户、启动次数、单次使用时长和留存率等指标。这些指标均已阐述过，此处不再赘述。以上提到的只是渠道质量评估的初步维度，如果还需要进一步研究渠道，尤其是研究到渠道防作弊层面，指标还需要更多，包括：判断用户使用行为是否正常的指标，如关键操作活跃量占总活跃的占比，用户激活 APP 的时间是否正

常；判断用户设备是否真实，如机型、操作系统等集中度的分析。

总之，如果要深入研究渠道作弊，算法的核心思想是研究推广渠道所带来的用户是否是真的“人”在用，从这个方向去设计相关的评估指标和算法，如某渠道带来的用户大部分集中在凌晨 2 点使用 APP，我们就认为这种渠道所带来的用户很可能不是正常人在使用，甚至可能是机器在作弊。

7.4.5　功能分析

功能分析主要分析功能活跃情况、页面访问路径以及转化率。这些指标需要功能运营的产品经理重点关注。

1. 功能活跃指标

功能活跃指标主要关注某功能的活跃人数、某功能新增用户数、某功能用户构成、某功能用户留存。这些指标的定义与 7.4.2 节介绍的指标类似。只是，本节只关注某一功能模块，而不是 APP 整体。

2. 页面访问路径分析

APP 页面访问路径统计用户从打开应用到离开应用整个过程钟每一步的页面访问和跳转情况。页面访问路径分析的目的是在达到 APP 商业目标之下帮助 APP 用户在使用 APP 的不同阶段完成任务，并且提高任务完成的效率。APP 页面访问路径分析需要考虑以下三方面问题：

(1) APP 用户身份的多样性，用户可能是你的会员或者潜在会员，有可能是你的同事或者竞争对手等；

(2) APP 用户目的多样性，不同用户使用 APP 的目的有所不同；

(3) APP 用户访问路径的多样性，即使是身份类似、使用目的类似，但访问路径也很可能不同。

因此，我们在做 APP 页面访问路径分析的时候，需要对 APP 用户做细分，然后再进行 APP 页面访问路径分析。最常用的细分方法是按照 APP 的使用目的来进行用户分类，如汽车 APP 的用户便可以细分为关注型、意向型、购买型用户，并对每类用户进行基于不同访问任务的路径分析，比如意向型的用户，他们进行不同车型的比较都有哪些路径，存在什么问题。还有一种方法是利用算法，基于用户所有访问路径进行聚类分析，基于访问路径的相似性对用户进行分类，再对每类用户进行分析。

3. 漏斗模型

漏斗模型是用于分析产品中关键路径的转化率，以确定产品流程的设计是否合理，分析用户体验问题。转化率是指进入下一页面的人数（或页面浏览量）与当前页面的人数（或页面浏览量）的比值。用户从刚进入到完成产品使用的某关键任务时（如购物），不同步骤之间的转换会发生损耗。如用户进入某电商网站，到浏览商品，到把商品放入购物车，最后到支付，每一个环节都有很多的用户流失损耗。

通过分析转化率，我们可以比较快定位用户使用产品的不同路径中，那一路径是否存在问题。当然，对于产品经理，其实不用每天都看转化率报表，我们可以对每天的转化率

进行连续性的监控，一旦转化率出现较大的波动，便发告警邮件给到相应的产品负责人，以及时发现产品问题。漏斗模型分析转化率如图 7.11 所示。

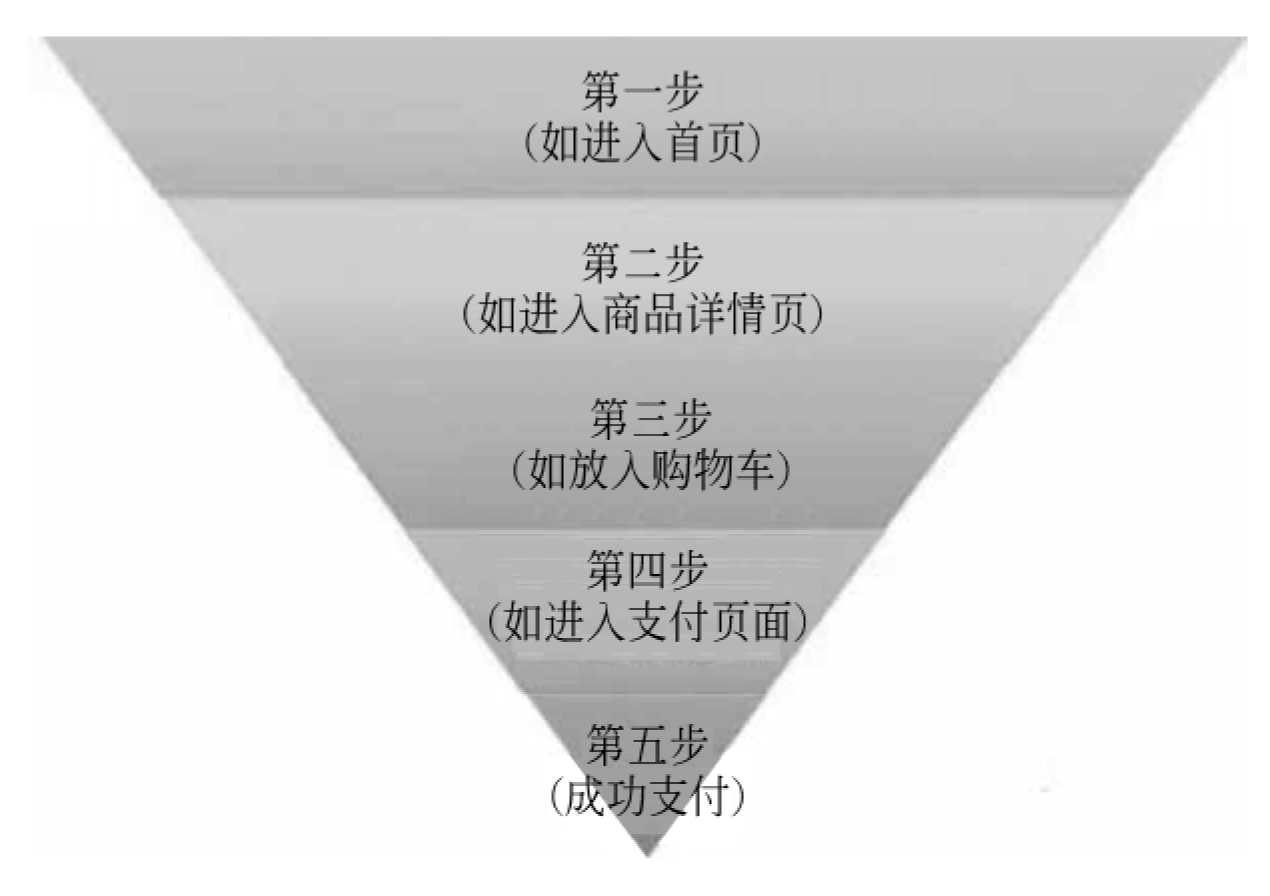


图 7.11　漏斗模型用于分析产品中关键路径的转化率

7.4.6　用户属性分析

用户属性分析主要从用户使用的设备终端、网络及运营商分析和用户画像角度进行分析。

1. 设备终端分析

设备终端的分析维度包括机型分析、分辨率分析和操作系统系统分析，在分析的时候，主要针对这些对象进行活跃用户、新增用户数、启动次数的分析。即分析不同机型的活跃用户数、新增用户数和启动次数，分析不同分辨率设备的活跃用户数、新增用户数和启动次数，分析不同操作系统设备的活跃用户数、新增用户数和启动次数。

2. 网络及运营商分析

网络及运营商主要分析用户联网方式和使用的电信运营商，主要针对这些对象进行活跃用户、新增用户数、启动次数的分析。即分析联网方式（包括 WiFi、2G、3G、4G）的活跃用户数、新增用户数和启动次数，分析不同运营商（中国移动、中国电信、中国联通等）的活跃用户数、新增用户数和启动次数。

3. 地域分析

主要分析不同区域，包括不同省市和国家的活跃用户数、新增用户数和启动次数。

4. 用户画像分析

用户画像分析包括人口统计学特征分析、用户个人兴趣分析、用户商业兴趣分析。人口统计学特征包括性别、年龄、学历、收入、支出、职业、行业等；用户个人兴趣指个人生活兴趣爱好的分析，如听音乐、看电影、健身、养宠物等；用户商业兴趣指房产、汽车、金融等消费领域的兴趣分析。用户画像这部分的数据需要进行相相关的画像数据采集，才可以支撑比较详细的画像分析。

7.5 大数据营销业务模型

7.5.1 大数据对业务模式的影响

大数据及其发挥的作用将影响到每一家公司——从财富500强企业到夫妻店——并从内到外地改变我们开展业务的方式。

公司在哪个领域运营，或者公司是什么规模，这都不要紧，因为数据收集、分析和解读变得更加轻松便捷，将从几个方面影响到每家公司。

1. 对所有公司来说，数据都将成为一项资产

如今，就连最小的公司也都在产生数据。如果公司有网站、有社交媒体账户、接受信用卡付款等，甚至哪怕它是一家只有一人经营的小店，都能从其客户、客户体验、网站流量等等方面收集数据。这意味着各种规模的公司都需要一个针对大数据的战略，并对如何收集、使用和保护数据制订计划。这也意味着精明的企业将开始向各公司提供数据服务，哪怕对方是一家非常小的公司。

它也意味着从来没想过大数据将"为它们所用的"企业和行业会争着迎头赶上。如果你拥有或经营一家企业，并且你想知道如何对企业做出改进，那么你需要借助数据，数据就是一项资产，它可用于改进企业运营情况。

2. 大数据能让公司收集更高质量的市场和客户情报

不管你喜不喜欢，你与之开展业务的公司了解你的很多情况——它们所掌握的有关你的信息的数量和类别每年都在扩大。每家公司（从监控我们开车情况的汽车制造商到了解我们打球频率和水平的网球拍生产商）都将对客户想要什么、使用什么、通常从哪个渠道购买等拥有更加深入的了解。

另一方面公司需要对制订和执行隐私政策采取积极主动的态度，所有的系统和安全防护措施都要到位，以保护这些用户数据。我们从近期免费升级的 Microsoft 10 身上可以看到，大多数人会允许公司收集这些数据，但他们希望公司对收集了什么数据以及为什么收集保持透明，同时他们希望可以选择不参与数据收集流程。

3. 大数据具备提高工作效率并改进运营的潜力

从使用传感器到追踪机器性能、优化送货路线、更好地追踪员工绩效甚至招募顶级人才，大数据具备能够提高几乎任何类型的企业及众多不同部门内部工作效率并改进运营的潜力。

公司可以使用传感器追踪货运和机器的运行情况，也可以追踪员工绩效。各公司已开始使用传感器追踪员工的移动、压力水平、健康状况甚至他们与谁交谈以及使用的语调等。

此外，如果数据能够成功量化一名优秀CEO所应具备的特质，它就能用来改进任何一个层级的人力资源和招聘流程。

数据正从IT部门脱离，成为一家公司中所有部门不可分割的一部分。

4. 数据可让公司改进客户体验并将大数据植入其提供的产品中

在所有可能的领域,公司都将使用它们收集的数据改进产品和客户体验。它不仅使用数据让自己的客户受益,还把数据作为一个新的产品提供给客户。

现代大型拖拉机公司所有新生产的拖拉机都配备了传感器,能够帮助该公司了解设备是如何使用的,同时预测并诊断故障。但公司安装传感器也是为了帮助农场主,为他们提供何时种植作物、在哪里种植、最佳的耕作和收割模式等等方面的数据。对于一家大型拖拉机公司来说,这已成为一个全新的收入来源。

随着我们生活中联网的事物越来越多——从智能恒温器到 Apple Watch 和健身追踪器——公司会有越来越多的数据、分析报告和信息回售给顾客。

7.5.2 大数据时代的网络化精确营销

营销策略制定的其中一大难题便是如何配置各项营销资源,在思考这个问题的时候,需要深入了解自家使用者的特性,并了解不同营销管道是否能与使用者的特性搭配。除此之外,分析现有营销管道的绩效,也是一项判断的重要依据。以下将介绍如何利用网站存取数据(Access Log,如 Google Analytics),初步分析各网络营销渠道的绩效。

要分析各营销渠道的绩效,首先需要定义绩效的指标,许多网站有其成立的目标,例如,销售商品、取得注册会员数等,这些指标在此统称为"转换数"(conversion)。定义绩效指标后,便可进行两个分析步骤:

(1) 统计各流量来源的转换数与转换率。

(2) 比较各流量来源的转换情况,拟定改善计划。

1. 步骤一:统计各流量来源的转换数与转换率

存取数据中有一项功能,能够追踪网站的流量来源,我们可以透过这项功能,将网站不同流量来源分类整理总流量、转换数及转换率。整理的同时,建议加入各流量来源的到达页面以及页面流程(称为一个沟通流程),更能交叉分析出有用的信息。

例如以关键词自然搜索为例,通过搜寻 beBit 这个公司名称进入网站的流量在过去一季共有 11 000 次,其中有 300 次成功注册会员(转换率为 3%)。搜寻 beBit 进入网站的沟通流程为:进站页面为首页,之后流经品类列表,最后到达商品页后成功转换(这是一个转换率为 3%,相对较好的沟通流程)。

2. 步骤二:比较各沟通流程的转换率,找出问题所在

搜集了各入口转换数的到达页及流入过程后(沟通流程),可以分析各个沟通流程与用户的沟通绩效。除了流程图之外,还可以作成表格整理流入过程,以便分析比较。

在站外广告与自然搜索的入口中,如果以商品页为到达页面,转换率明显偏低。此时我们可以回头检视是不是站外广告与商品页无法连贯说服使用者。在这里还可以更进一步进行交叉分析,站外广告进入商品页的用户特征(例如,重复造访 vs 新造访、会员 vs 非会员、人口变量……),取得该营销渠道无法成功转换使用者的更深度因素,以判别是要改善该渠道的沟通内容,或者是舍弃该营销渠道。

以上是从分析现有的营销渠道绩效,做出营销资源规划的初步判断。倘若需要评估

新的营销渠道的投资潜力，建议也是回到使用者角度，了解网站的目标用户接触到该渠道的情境（时间、地点、方式、心态），以判断该渠道与使用者的接触点与说服力，进一步判定是否具有投资潜力。

7.5.3　移动互联和大数据时代的电子商务

我们有幸生活在一个互联网时代，尤其是移动互联的时代，这是一个大数据的时代。这个时代里人们的生活方式正在被改变，创新的商务模式不断地涌现。

中国的电子商务已经全面超过美国，不管是从线上的总销售额，还是线上销售在全社会零售的占比，还是增速。中国的网购人群已经超过美国人口，但是发展的空间还是巨大的。美国的网购渗透率超过了75%，而中国的网购渗透率才刚过50%，潜力巨大。

在这个过程中，移动商务的发展是井喷式的，移动商务很快就成了所有电商的主战场。

电子商务有以下一些优势：

（1）首先它不受地域限制，一网覆盖全国乃至全球。第二，它不受时间限制，它可以7×24小时服务。第三，它可以有无穷的货架，增加商品只是增加服务器。还有一个就是大数据。大数据允许我们更多地了解顾客，提供精准的营销和个性化服务。而移动商务在这几个优势的基础上又增加了很多的新的优势。

（2）移动客户端，尤其是智能手机的发展，其扫描和图像识别功能可以方便顾客的搜索。现在已经可以用智能手机拍摄一个图片，把这个图片里面的所有商品识别出来，把这个商品和其最契合的款式、颜色和品牌，找出来，匹配起来，再迅速地链接到相关的店里去，方便顾客立即购买。

有了智能手机，可以随时知道你在什么地方，及时告诉你周边有什么服务有什么商品适合你。有了大数据可以分析到这个顾客的喜好，可以分析跟这个顾客同类画像的顾客群的喜好，给你推送适合你的商品。十多年前，大家谈物联网，可是如果没有智能手机，没有可穿戴设备，没有各种感应设备的话，也只是个概念。现在这些概念都变成现实。

移动购物有新的特征，因为大家把零散的碎片化的时间利用起来，可以在上班的路上，可以在地铁里，可以在公交车上，可以在旅游的过程中，可以在任何的场景随时购物。所以购物的特征更频繁、更零碎，每单总价降低，但是购买的频次增高。大家发现，回到家的晚上，甚至躺在床上都可以购物，节假日根本不需要打开电脑，随时可以购物。这是新的特征。

（3）大众营销即将消失，至少这个时代取代它的是窄众营销。现在，我们将顾客分为宅男、丽人、辣妈、新客四个角色，这样至少可以部分精准地为顾客服务。手机客户端现在能做到千人四面，我们更希望做到千人千面。但是终极目标是精准营销，每一个顾客都有适合自己的最精准的信息，比如说我上新浪体育只看NBA或高尔夫，我不喜欢看足球看体操，每天给我看这些没有用，给我看NBA就行了。我是高血压患者，不需要看糖尿病的内容，你向我推广糖尿病的药是没有任何意义的。这时候就节约了营销成本，最后也是让顾客受益。生产也从早期的批量生产变成批量定制，终极目标是C2B，针对每一个个人的喜好和所需来制造。

(4) 电商是更智能化更本地化、社交化和个性化的。大家设想一下，如果上任何一个网站也好 APP 也好，一上去就知道你是谁，知道你的画像，知道你在什么地方，知道现在是什么季节，知道在这个季节里面适合你的是什么样的服务和商品，最后给你提供你最适合的需求。你感觉这个网站就是为你服务的，这就是未来的电商，这就是基于移动和基于大数据的电商。

7.5.4 大数据营销的定义与特点

大数据营销是基于多平台的大量数据，依托大数据技术的基础上，应用于互联网广告行业的营销方式。大数据营销衍生于互联网行业，又作用于互联网行业。依托多平台的大数据采集，以及大数据技术的分析与预测能力，能够使广告更加精准有效，给品牌企业带来更高的投资回报率。

1. 大数据营销的定义

大数据营销是指通过互联网采集大量的行为数据，首先帮助广告主找出目标受众，以此对广告投放的内容、时间、形式等进行预判与调配，并最终完成广告投放的营销过程。

大数据营销，随着数字生活空间的普及，全球的信息总量正呈现爆炸式增长。基于这个趋势的，是大数据、云计算等新概念和新范式的广泛兴起，它们无疑正引领着新一轮的互联网风潮。

2. 大数据营销的特点

1) 多平台化数据采集

大数据的数据来源通常是多样化的，多平台化的数据采集能使对网民行为的刻画更加全面而准确。多平台采集可包含互联网、移动互联网、广电网、智能电视未来还有户外智能屏等数据。

2) 强调时效性

在网络时代，网民的消费行为和购买方式极易在短的时间内发生变化。在网民需求点最高时及时进行营销非常重要。全球领先的大数据营销企业 AdTime 对此提出了时间营销策略，它可通过技术手段充分了解网民的需求，并及时响应每一个网民当前的需求，让他在决定购买的“黄金时间”内及时接收到商品广告。

3) 个性化营销

在网络时代，广告主的营销理念已从“媒体导向”向“受众导向”转变。以往的营销活动须以媒体为导向，选择知名度高、浏览量大的媒体进行投放。如今，广告主完全以受众为导向进行广告营销，因为大数据技术可让他们知晓目标受众身处何方，关注着什么位置的什么屏幕。大数据技术可以做到当不同用户关注同一媒体的相同界面时，广告内容有所不同，大数据营销实现了对网民的个性化营销。

4) 性价比高

和传统广告“一半的广告费被浪费掉”相比，大数据营销在最大程度上让广告主的投放做到有的放矢，并可根据实时性的效果反馈，及时对投放策略进行调整。

5）关联性

大数据营销的一个重要特点在于网民关注的广告与广告之间的关联性，由于大数据在采集过程中可快速得知目标受众关注的内容，以及可知晓网民身在何处，这些有价信息可让广告的投放过程产生前所未有的关联性。即网民所看到的上一条广告可与下一条广告进行深度互动。

3. 大数据营销的实现过程

大数据营销并非是一个停留在概念上的名词，而是一个通过大量运算基础上的技术实现过程。事实上，国内的很多以技术为驱动力的企业也在大数据领域深耕不辍。

全球领先的大数据营销平台 AdTime 率先推出了大数据广告运营平台——云图。据介绍，云图的云代表云计算，图代表可视化。云图的含义是将云计算可视化，让大数据营销的过程不再神秘。云图是 AdTime 构建的大数据平台系统，该系统具备海量数据、实时计算、跨网络平台汇聚、多用户行为分析、多行业报告分析等特点。

大数据营销是基于大数据分析的基础上，描绘、预测、分析、指引消费者行为，从而帮助企业制定有针对性的商业策略。

大数据营销中所依赖的数据，往往是基于 Hadoop 架构分类的静态人群属性和兴趣爱好常量，这导致了大数据营销在本质上很难去控制和捕获用户的需求。

4. 契机

第一，用户行为与特征分析。

只有积累足够的用户数据，才能分析出用户的喜好与购买习惯，甚至做到“比用户更了解用户自己”。这一点，才是许多大数据营销的前提与出发点。

第二，精准营销信息推送支撑。

精准营销总在被提及，但是真正做到的少之又少，反而是垃圾信息泛滥。究其原因，主要就是过去名义上的精准营销并不怎么精准，因为其缺少用户特征数据支撑及详细准确的分析。

第三，引导产品及营销活动投用户所好。

如果能在产品生产之前了解潜在用户的主要特征，以及他们对产品的期待，那么你的产品生产即可投其所好。

第四，竞争对手监测与品牌传播。

竞争对手在干什么是许多企业想了解的，即使对方不会告诉你，但你却可以通过大数据监测分析得知。品牌传播的有效性亦可通过大数据分析找准方向。例如，可以进行传播趋势分析、内容特征分析、互动用户分析、正负情绪分类、口碑品类分析、产品属性分布等，可以通过监测掌握竞争对手传播态势，并可以参考行业标杆用户策划，根据用户声音策划内容，甚至可以评估微博矩阵运营效果。

第五，品牌危机监测及管理支持。

新媒体时代，品牌危机使许多企业谈虎色变，然而大数据可以让企业提前有所洞悉。在危机爆发过程中，最需要的是跟踪危机传播趋势，识别重要参与人员，方便快速应对。大数据可以采集负面定义内容，及时启动危机跟踪和报警，按照人群社会属性分析，聚类

事件过程中的观点，识别关键人物及传播路径，进而可以保护企业、产品的声誉，抓住源头和关键结点，快速有效地处理危机。

第六，企业重点客户筛选。

许多企业家纠结的事是：在企业的用户、好友与粉丝中，哪些是最有价值的用户？有了大数据，或许这一切都可以更加有事实支撑。从用户访问的各种网站可判断其最近关心的东西是否与你的企业相关；从用户在社会化媒体上所发布的各类内容及与他人互动的内容中，可以找出千丝万缕的信息，利用某种规则关联及综合起来，就可以帮助企业筛选重点的目标用户。

第七，大数据用于改善用户体验。

要改善用户体验，关键在于真正了解用户及他们所使用的你的产品的状况，做最适时的提醒。例如，在大数据时代或许你正驾驶的汽车可提前救你一命。只要通过遍布全车的传感器收集车辆运行信息，在你的汽车关键部件发生问题之前，就会提前向你或 4S 店预警，这决不仅仅是节省金钱，而且对保护生命大有裨益。事实上，美国的 UPS 快递公司早在 2000 年就利用这种基于大数据的预测性分析系统来检测全美 60 000 辆车辆的实时车况，以便及时地进行防御性修理。

第八，SCRM 中的客户分级管理支持。

面对日新月异的新媒体，许多企业通过对粉丝的公开内容和互动记录分析，将粉丝转化为潜在用户，激活社会化资产价值，并对潜在用户进行多个维度的画像。大数据可以分析活跃粉丝的互动内容，设定消费者画像各种规则，关联潜在用户与会员数据，关联潜在用户与客服数据，筛选目标群体做精准营销，进而可以使传统客户关系管理结合社会化数据，丰富用户不同维度的标签，并可动态更新消费者生命周期数据，保持信息新鲜有效。

第九，发现新市场与新趋势。

基于大数据的分析与预测，对于企业家提供洞察新市场与把握经济走向都是极大的支持。

第十，市场预测与决策分析支持。

对于数据对市场预测及决策分析的支持，过去早就在数据分析与数据挖掘盛行的年代被提出过。沃尔玛著名的“啤酒与尿布”案例即是那时的杰作。只是由于大数据时代上述 Volume(规模大)及 Variety(类型多)对数据分析与数据挖掘提出了新要求。更全面、更及时的大数据，必然对市场预测及决策分析进一步上台阶提供更好的支撑；似是而非或错误的、过时的数据对决策者是灾难。

7.5.5 网络营销大数据实际操作

对很多企业来说，大数据的概念已不陌生，但如何在营销中应用大数据仍是说易行难。其实，作为大数据最先落地也最先体现出价值的应用领域，网络营销的数据化之路已有成熟的经验及操作模式。

1. 获取全网用户数据

首先需要明确的是，仅有企业数据，即使规模再大，也只是孤岛数据。在收集、打通企

业内部的用户数据时，还要与互联网数据统合，才能准确掌握用户在站内站外的全方位的行为，使数据在营销中体现应有的价值。在数据采集阶段，建议在搜集自身各方面数据形成 DMP 数据平台后，还要与第三方公用 DMP 数据对接，获取更多的目标人群数据，形成基于全网的数据管理系统。

2. 让数据看得懂

采集来的原始数据难以懂读，因此还需要进行集中化、结构化、标准化处理，让“天书”变成能看得懂的信息。

这个过程中，需要建立、应用各类“库”，如行业知识库（包括产品知识库、关键词库、域名知识库、内容知识库）；基于“数据格式化处理库”衍生出来的底层库（用户行为库、URL 标签库）；中层库（用户标签库、流量统计、舆情评估）；用户共性库等。

通过多维的用户标签识别用户的基本属性特征、偏好、兴趣特征和商业价值特征。

3. 分析用户特征及偏好

将第一方标签与第三方标签相结合，按不同的评估维度和模型算法，通过聚类方式将具有相同特征的用户划分成不同属性的用户族群，对用户的静态信息（性别、年龄、职业、学历、关联人群、生活习性等）、动态信息（资讯偏好、娱乐偏好、健康状况、商品偏好等）、实时信息（地理位置、相关事件、相关服务、相关消费、相关动作）分别描述，形成网站用户分群画像系统。

4. 制定渠道和创意策略

根据对目标群体的特征测量和分析结果，在营销计划实施前，对营销投放策略进行评估和优化。如选择更适合的用户群体，匹配适当的媒体，制定性价比及效率更高的渠道组合，根据用户特征制定内容策略，从而提高目标用户人群的转化率。

5. 提升营销效率

在投放过程中，仍需不断回收、分析数据，并利用统计系统对不同渠道的类型、时段、地域、位置等价值进行分析，对用户转化率的贡献程度进行评估，在营销过程中进行实时策略调整。

对渠道依存关系进行分析：分析推广渠道的构成类型与网站频道、栏目的关联程度（路径图形化＋表格展示）；

对流量来源进行分析：分析网站各种推广渠道类型的对网站流量的贡献程度；

对用户特征及用户转化进行分析：分析各个类型的推广渠道所带来的用户特征、各个推广渠道类型转化效率、效果和 ROI。

6. 营销效果评估、管理

利用渠道管理和宣传制作工具，利用数据进行可视化的品牌宣传、事件传播和产品，制作数据图形化工具，自动生成特定的市场宣传报告，对特定宣传目的报告进行管理。

7. 创建精准投放系统

对于有意领先精准营销的企业来说，则可更进一步，整合内部数据资源，补充第三方

站外数据资源，进而建立广告精准投放系统，对营销全程进行精细管理。

7.5.6 数据营销方法论

Google 每天要处理大约 24PB 的数据，Facebook 每天要处理 23TB 的数据，Twitter 每天处理 7TB 的数据，百度每天大概新增 10TB 的数据。

腾讯每日新增加 200～300TB 的数据，淘宝每日订单超过 1000 万，阿里巴巴已经积累的数据量超过 100 个 PB。考虑一下，为什么越是行业垄断巨头就越拥有海量数据呢？

对任何拥有特有数据的公司，都应该考虑怎么让数据盈利。

1. 数据收集没想象中那么复杂，重要的是发现

很多企业甚至是互联网企业，或者不知道该如何使用手中已有的数据资源，白白浪费掉优化改进的好机会；或者认为大数据只有 BAT 这样的互联网巨头才有，一个小网站或 APP 应用是没有大数据的，果真是如此吗？

看一个简单的例子——微博段子手们最平常不过的数据收集。

抛出一个限定话题得到各方粉丝回应，第二天可参照由微博点赞自动生成具有代表性的意见进行概括归纳，将 1k＋的评论总结起来制成 9 条 Tips，二次加工后发出获得 6k＋转发、4k＋评论和 4k＋赞。

一个网站或一个 APP 所包含的数据信息都是数字营销的基础。

通过分析来自网站及竞争对手的定性与定量数据，可以驱动用户及潜在用户在线体验的持续提升，并提高数字营销业绩，如图 7.12 所示。

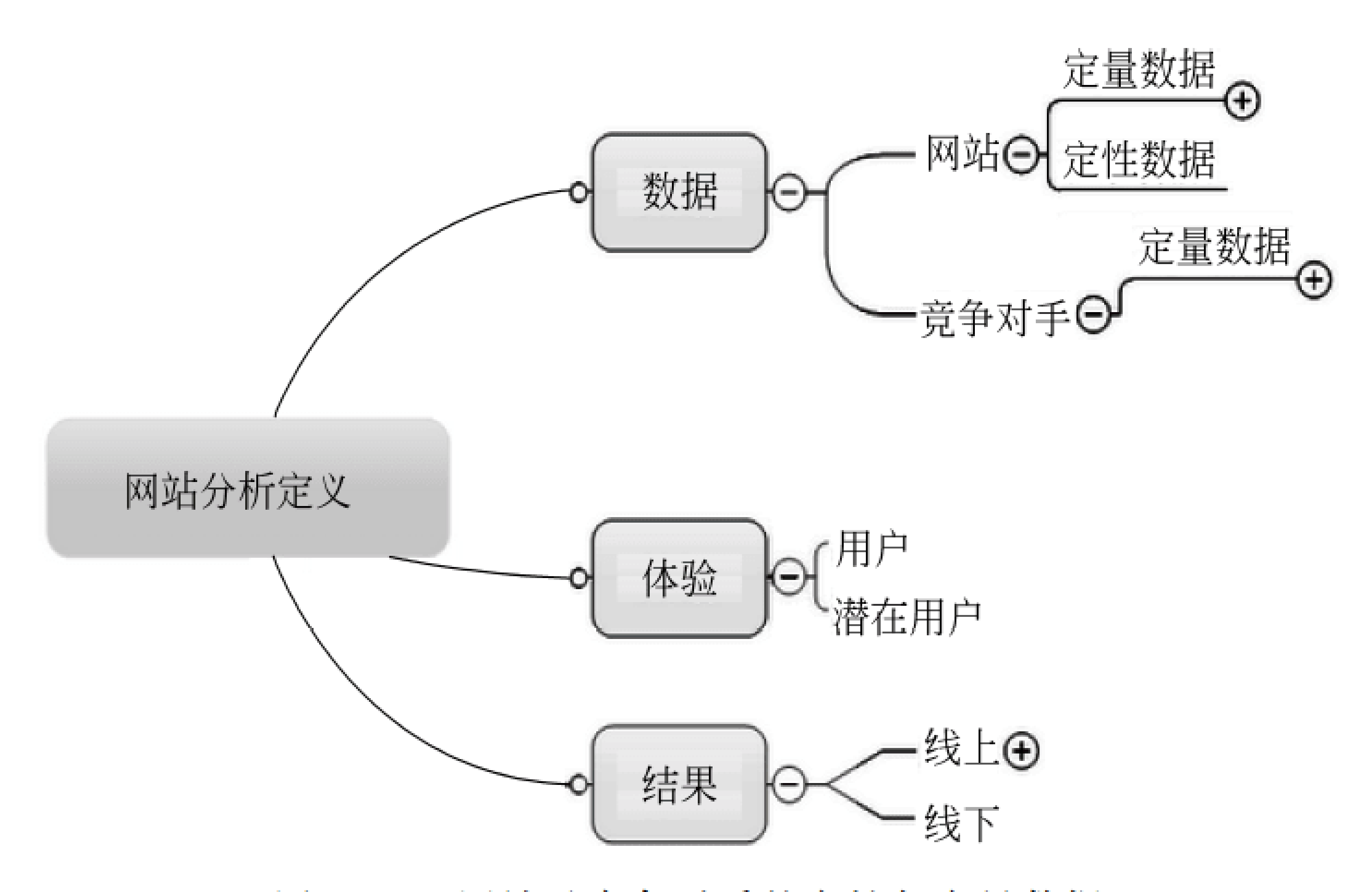


图 7.12 网站及竞争对手的定性与定量数据

又如，法国的一些航空公司推出免费的 APP 方便旅客在移动设备上跟踪自己的行李，之后在追踪的数据平台上发现一部分商务旅行客户中途在某一城市进行短暂的商业会晤不需入住酒店，行李成了累赘，于是航空公司推出专人看管全程可追踪的增值服务，此项服务每周的新增价值大概可达 100 万美元。

正是基于对数据的洞察产出附加价值。对数据的掌控，就是对市场的支配，意味着丰厚的投资回报。

2. 数据是有情绪的，假如别人要你推荐一只股票

数据的形式多种多样，呈数量级爆发的 UGC 内容可以被我们拿来运用？一个新颖点的例子，譬如对于从 5100 点＋飞泻而下的中国股市，股民巨量的埋怨和牢骚能以怎样的数据化形式展示？

“除了耐心等待，最好再找个地方让自己发泄一下，找些跟自己同病相怜的人，还能缓解下压力，避免跳楼。弹幕，就是最好的形式了。”一旦有人建了一个网站，在 K 线图上配上弹幕供吐槽……

结果被同样郁闷的股民汇集出的数据随着 K 线走势变化拥有了实时鲜明的情绪特征，可以在一定程度预估使用者下一步卖出或继续持有的动向，如图 7.13 所示。

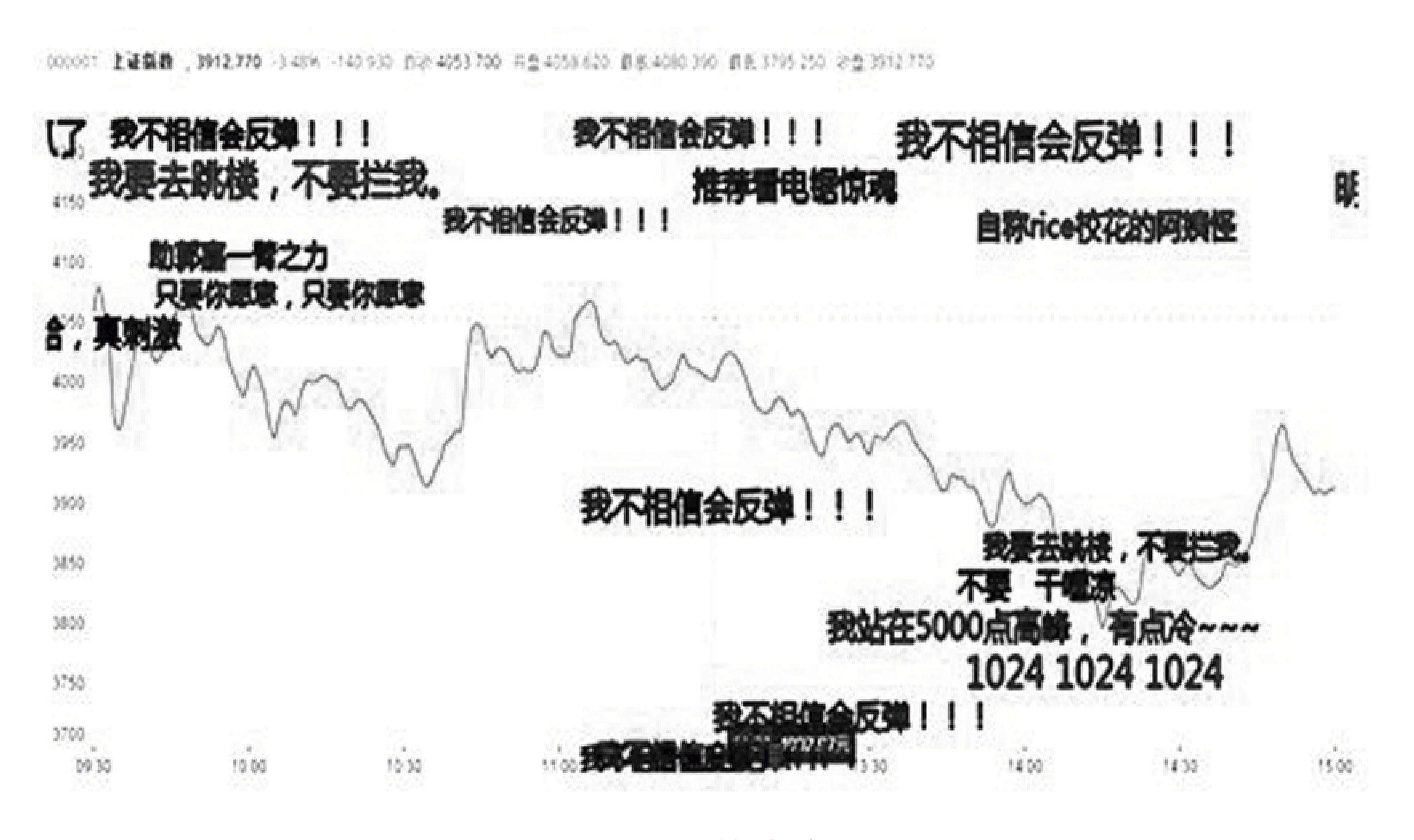


图 7.13　股民情绪特征预测

1）股市数据

拿买股票来说，推荐者会继续购买并且推荐给其他人来加速某个公司股票(或实际产品)的成长；而贬损者则能破坏其名声，不仅仅停止购买，而且劝说周围朋友，在负面的口碑中阻止其成长，NPS 净推荐值则反映了类似多与空、褒与贬这两股力量较量的结果。

回到广告，这些来源于门户或垂直类网站、电商平台购物用户的打分与评论、社会化媒体如微博、论坛、微信、应用等的用户评论文本数据以及客服系统的语音数据和评价文本数据，可以统称为“用户反馈数据”，如图 7.14 所示。

2）用户调查表

我们可以结构化处理后，进行数据挖掘，识别“贬损者”和“推荐者”，全面和快速地计算 NPS，并了解“贬损者”的贬损原因。

若进一步关联整合“用户行为数据”，我们还可以了解“贬损者”的历史“用户行为数据”，有利于更好地洞察用户情况，优化用户体验和改进产品方向；同时还能定向向“推荐者”展开更多的优惠促销或附加增值服务。

当广告商掌握了数据，能够向客户传输更加相关的和更加有趣的信息，潜在客户们甚至可以根据自己的需求定制一些广告信息，可能会做出更好的购物决策，并有助于广告商

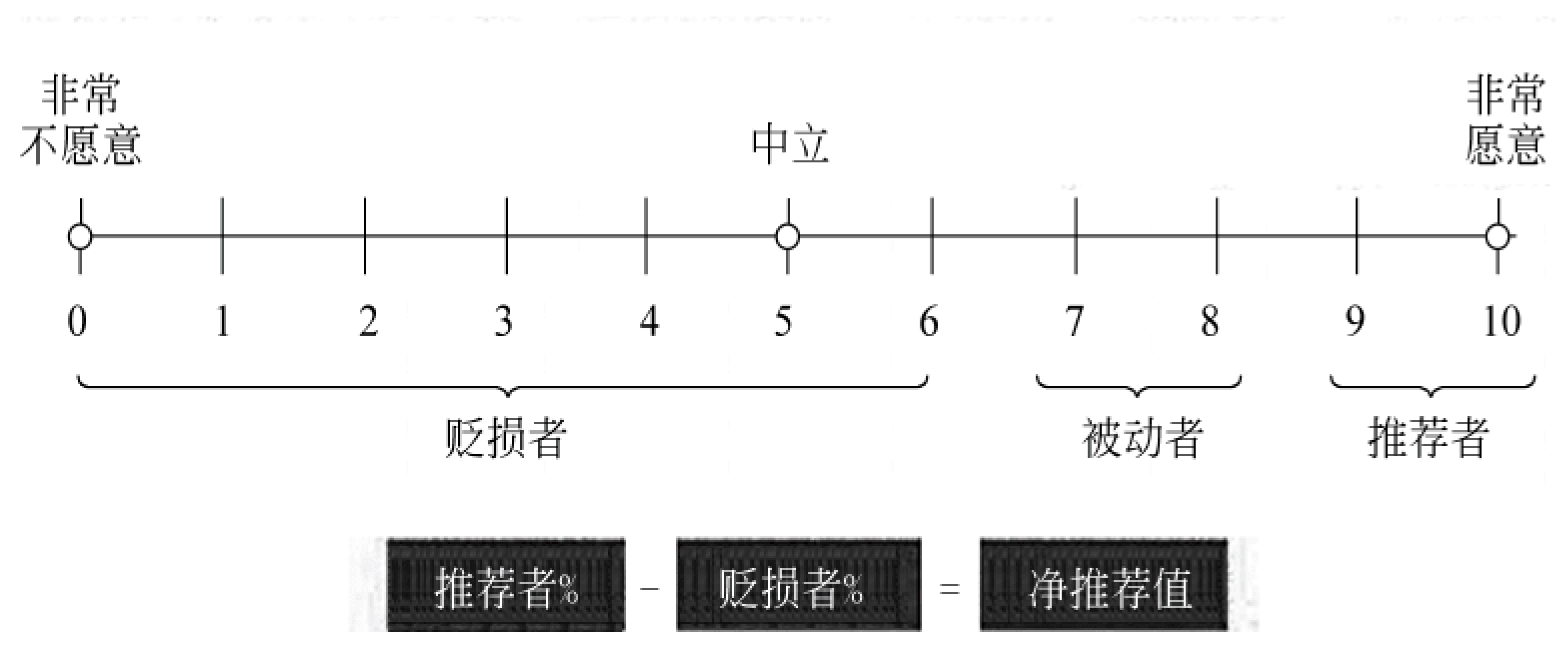


图 7.14 用户反馈数据

提升销售业绩。

3. 基本的5W1H问答也能玩转消费行为数据(科特勒(Kotler)行为选择模型范例)

科特勒(Kotler 行为选择)模型从市场的特点来探讨消费者行为,更容易进行定量研究。

以推广营销某款手机为例,我们将要研究的数据可综合为5W1H:

(1) Who & Whom:购买这款手机的人群分类?还要弄清谁是决策者,谁是使用者,谁对决定购买有重大影响以及谁是实际购买者;

(2) What:不同手机品牌的市场占有率、具体型号的销售情况;

(3) When:了解在具体的季节、时间甚至时点所发生的购买行为,比如配合节假日促销;

(4) Where:研究适当的销售渠道和地点,还可以进一步了解消费者是在什么样的地理环境、气候条件甚至于地点场合使用手机;

(5) How:了解消费者怎样购买、喜欢什么样的促销方式,比如是去线下体验店还是看测评视频等;

(6) Why:探索消费者行为动机和偏好,比如为什么喜欢特定款手机并拒绝别的品牌或型号?

不同特征的消费者会产生不同的心理活动的过程,通过其决策过程导致了一定的购买决定,最终形成了消费者对产品、品牌、经销商、购买时机、购买数量的选择,如图 7.15 所示。

数字营销人员如果能比较清楚地了解各类购买者对不同形式的产品、服务、价格、促销方式的真实反应,就能够适当地影响、刺激或诱发购买者的购买行为。数据的应用可以贯穿营销价值链的广告、公关、官网、电商、CRM 各个环节,覆盖用户能力会更加全面和强大。

4. 数据是拿来用的,不仅仅是拿来看

买一只股票尚需数据分析,展开一项持续的广告营销活动当然更应该建立在有数据衡量的基础上。

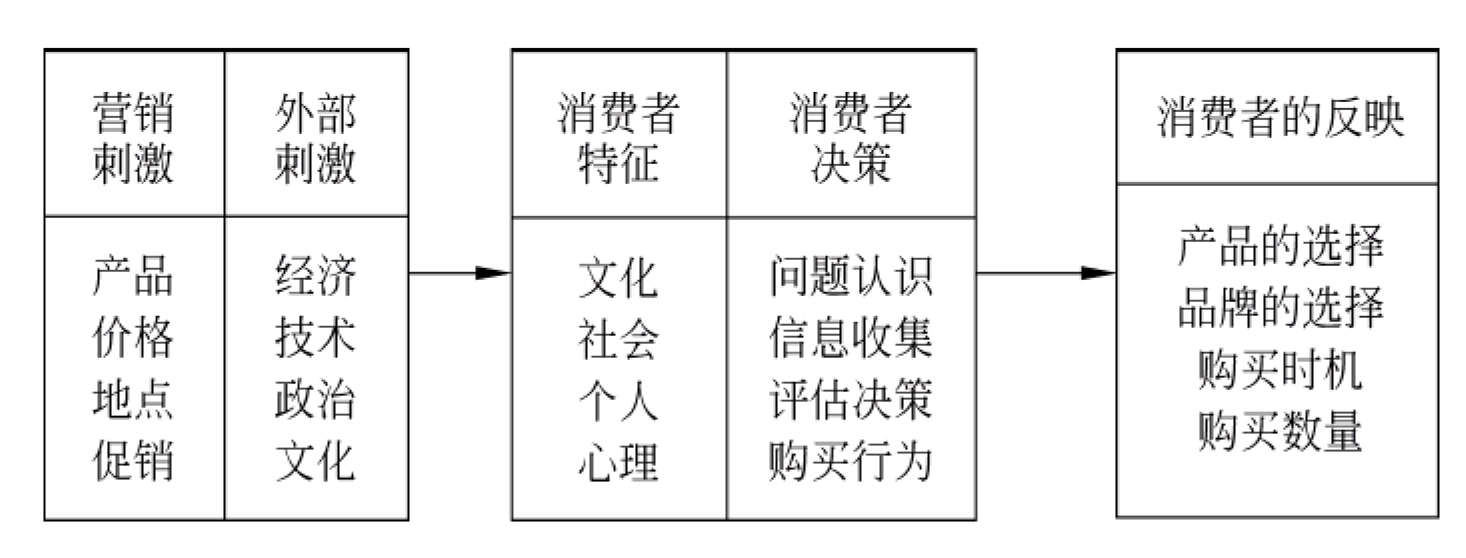


图 7.15　科特勒行为选择模型

比如 Uber 的数据科学家建立了“基于地理位置的打车需求模型”(Location-based demand model),每天实时更新的热点地图可以有效帮助车主缩短空载时间,同时帮乘客减少等待时长。

PRADA 在纽约的旗舰店中每件衣服上都有 RFID 码,每一件衣服在哪个旗舰店什么时间被拿进试衣间停留多长时间,数据都被存储起来加以分析。某一系列衣服销量很低,以往是被直接“干掉”。但如果 RFID 传回的数据显示这系列的衣服虽然销量低但进试衣间的次数多,那就能另外说明一些问题。

也许在某个细节的微小改变就会重新创造出一件非常流行的产品,这类衣服的下场会截然不同。有点像电商分析购物车数据来提高转化率,若大量客户都选中了某件商品放入购物车却没有最终结算,说明它是热门产品,但可能有些小问题,适当变更价格或服务条款可能就会产生巨大的变化。

数据的使用能够使对企业的经营对象从客户的粗略归纳还原成一个个活生生的客户,了解他们喜欢什么讨厌什么,并更有针对性,越能满足客户的需要,ROI 就更高。

广告主通过数字营销,更可能运用全新的视角来发现新的商业机会和重构新的商业模式。过去看不到的东西都能看到了,即有了全新的视野。

7.6　基于社会媒体的分析预测技术

7.6.1　基于空间大数据的社会感知

大数据时代产生了大量具有时空标记、能够描述个体行为的空间大数据,如手机数据、出租车数据、社交媒体数据等。这些数据为人们进一步定量理解社会经济环境提供了一种新的手段。近年来,计算机科学、地理学和复杂性科学领域的学者基于不同类型数据开展了大量研究,试图发现海量群体的时空行为模式,并建立合适的解释性模型。

“社会感知”(social sensing)就是借助于各类空间大数据研究人类时空间行为特征,揭示社会经济现象的时空分布、联系及过程的理论和方法。值得一提的是,与强调基于多种传感设备采集微观个体行为数据的社会感知计算(socially aware computing)相比,社会感知更加强调群体行为模式以及背后地理空间规律挖掘。

社会感知数据可从三个方面获取人的时空间行为特征:

(1) 对地理环境的情感和认知,如基于社交媒体数据获取人们对于一个场所的感受;

(2) 在地理空间中的活动和移动,如基于出租车、签到等数据获取海量移动轨迹;

(3) 个体之间的社交关系,如基于手机数据获取用户之间的通话联系信息。由于空间大数据包含了海量人群的时空间行为信息,使得我们可以基于群体的行为特征揭示空间要素的分布格局、空间单元之间的交互以及场所情感与语义。

空间大数据提供的社会感知手段,为地理学乃至相关人文社会科学研究开启了一种“由人及地”的研究范式。而“社会感知”这一概念,正是概括描述了空间大数据在相关研究与应用中所提供的数据以及方法上的支撑能力。

1. 社会感知分析方法

根据社会感知的概念,对于空间大数据的研究可以分为“人”和“地”两个层面。前者关注人的空间行为模式,以及模式所受到的地理影响;后者则侧重于在群体行为模式的基础上,探讨地理环境的相关特征。

2. 个体行为模式分析法

空间大数据可以感知人的三个方面的空间行为模式,如图 7.16 所示。其中,移动是个体层次空间行为最直接的外在表现。由于大数据对于移动轨迹的获取能力较强,因此目前的研究多集中在移动模式和模型的建立。

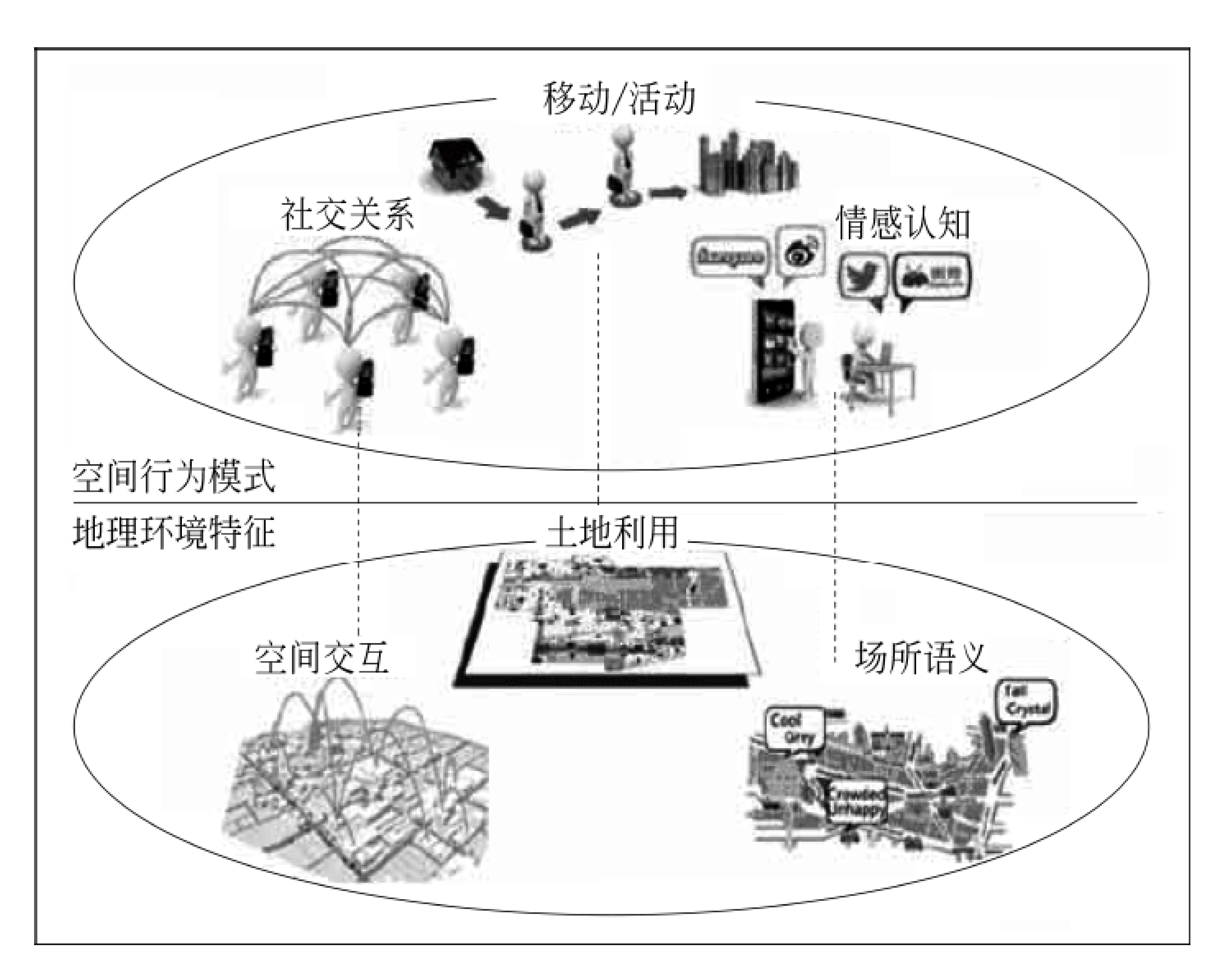


图 7.16 社会感知研究框架

动物以及人在空间中移动所展示的规律性是复杂系统领域研究的一个重要议题。每个个体的移动模式可以表示为随机游走(random walk)模型。通过对动物的移动进行观察,发现其移动步长和角度的统计分布特征呈现一定的模式,提高了觅食的效率。当移动方向均匀分布,而步长为幂律分布,且指数在 1~3 之间时,移动为列维飞行模型(Levy flight),如图 7.17 所示。

与动物相比,人的出行目的更加多样化,并且存在一个或者多个频繁重访地点,这使得人的移动模式与动物的移动模式存在机理上的差异。在海量个体移动轨迹数据的支持下,我们可以观察人的移动模式并构建相应的解释模型。从布罗克曼(Brockmann)等人

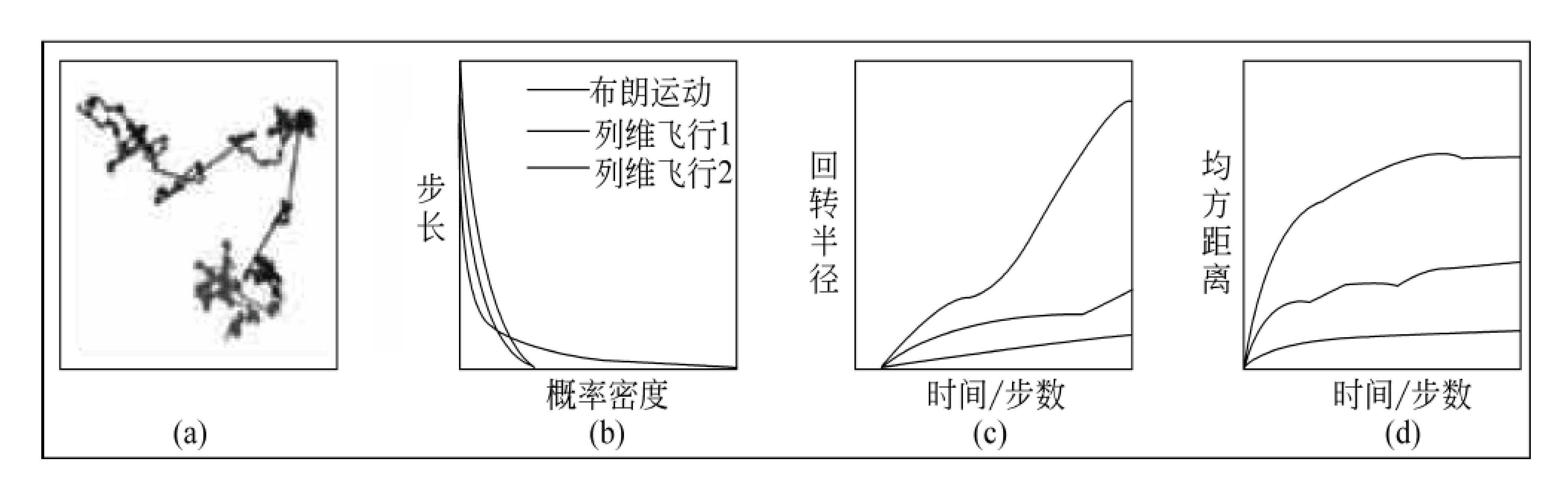


图 7.17　列微飞行模型的移动步长分布以及扩展特征

发表在《自然》上的基于钱币追踪数据开展的研究开始，许多学者利用手机、出租车、社交媒体签到等数据探讨了人的移动模式，并且试图建立解释性模型。

步长的统计分布是移动性模式表达中的重要元素。对于移动轨迹而言，由于距离衰减，使得长距离出行的概率较低，而短距离出行的概率较高。表征这种分布特征的函数有幂律分布、指数分布、指数截断的幂律分布等。许多学者试图建立模型以解释观察到的人类移动模式。除了距离衰减影响外，解释移动模式需要考虑的因素还包括地理环境和个体的空间行为特征。其中地理环境因素决定了潜在的个体移动到访点的空间分布，该分布通常与人口密度分布正相关；而个体的空间行为特征则反映了人们移动中的一些个性化的规律。

目前得到较多关注的是个体轨迹中的重访点，这是人类移动和动物移动存在较大差异的方面。人类移动存在家和工作地等频繁重访的地点，具有较高的可预测性。在地理环境分布特征方面，我们通常从城市范围内及城市间两个尺度分别探讨移动性模式。城市范围内的移动受到城市用地结构的影响。

对于一个城市而言，通常市中心区土地开发强度较大，居民出行的密度相对较高，而在城市边缘地区，土地利用强度和出行密度都相对较低。这种地理环境分布模式使得城市尺度的移动步长分布尾部不那么“重”。而对于城市间的移动，城市体系中不同规模的城市空间分布同样影响了观测到的移动模式。

目前研究所采用的空间大数据多数都是“移动轨迹丰富，活动信息不足”，这使得轨迹背后丰富的语义信息（尤其是出行目的信息）缺失。在交通地理学研究中，出行目的是理解出行移动模式的基础，不同的出行目的受到空间的约束也不同。一些学者试图结合轨迹数据、时间约束以及地理环境特征，推断出行目的，从而达到充实轨迹语义的目的。

个体层次的时空间行为除了移动和活动外，社交关系（social ties）也是很重要的要素。利用空间大数据可以揭示社交关系背后的地理影响。这方面的研究主要包括个体地理位置对于个体间社交关系的影响以及个体空间移动与社交关系的相互作用两个方向，目的是探求空间距离和时空共现（spatio-temporal co-occurrence）与社交关系之间的量化联系。

3. 活动时间变化特征分类法

不同类型的大数据可以揭示一个区域或城市的活动以及人口分布状态。大数据的时间标记可以用于解释人口分布的动态变化特征。这种变化特征往往具有较强的周期性。

对于城市研究而言，尤其以日周期变化最为明显。城市居民在居住地点和工作地点之间的通勤行为产生了相关地理单元人口密度的时变特征。因此，我们可以基于城市不同区域对应的活动日变化曲线来研究其用地特征和在城市运行中所承载的功能。

利用空间大数据所提取的活动分布特征感知土地利用类别的基本依据是活动量日变化特征对地块的指示能力。提取特征时通常采用非监督分类方法，最常用的算法有 *k*-平均算法(*k*-mean)聚类、*k*-中心点算法(*k*-medoid)聚类等。我们经常可以看到相同的土地覆被对应不同的居民活动特征，而外形相近的建筑可能承担了不同的社会功能，与之相较，利用大数据提取活动分布特征的方法从活动角度更为全面地解读了城市土地利用情况。

在分类过程中，因为功能相同的地块存在活动强度的差异，如高密度居民区和低密度居民区，尽管人口总量不同，但是其人口密度日变化特征相似，故而在非监督分类过程中，通常需要对活动时变曲线进行归一化处理。

此外，考虑城市居民工作日和周末的不同活动特征，在一些研究中，会将工作日数据和非工作日数据分开处理。由于空间大数据所提取的活动时空分布信息可以处理成与传统遥感数据相似的形式，因此除了非监督分类外，一些图像处理方法也可以应用于社会感知数据。

近年来，也有一些研究采用主成分分析以及非负矩阵分解方法，识别一个城市不同区域活动变化的全局和局部变化特征。此外，张量(tensor)也是分析时空大数据的有效工具，张量模型的高阶(high order)表达能力能够描述时空数据在时间、空间、个体状态等多方面的特征。

4. 场所情感及语义分析法

社交媒体(推特、微博等)中包含了大量文本数据，成为语义信息获取的重要来源。带有位置的社交媒体数据通常占 3%，研究者可以利用这部分数据揭示与地理位置有关的语义信息。目前的研究主要包括三个方向：

(1) 获取一个场所的主题词；

(2) 获取与场所有关的情感信息，如高兴还是抑郁；

(3) 获取对于特定事件(如灾害、事故、疾病)的响应。由于社交媒体数据是大量用户自发创建的，分析语义信息及其时空模式有助于政策制定者了解社情民意并制定相关公共政策。在社交媒体文本语义处理中，潜在狄利克雷分配(Latent Dirichlet Allocation，LDA)模型被广泛应用，以确定每条信息所表示的主题以及相关的情绪信息。然而，由于社交媒体数据中每条文本存在字数的限制，并且内容随意性较强，因此如何从中挖掘更加精确的、有意义的信息，尚需进一步研究。

近年来，深度学习技术的发展使得自动提取识别照片语义信息成为可能。一些研究基于对照片共享网站带有时空标记的图像进行内容分析，揭示地理环境的特征。

与基于文本的语义信息提取相比，照片语义信息更为客观且丰富。每张照片反映了拍照者对于场所的感知。考虑到文本和照片不同的表达能力，可以认为结合文本和照片语义信息，能够全面捕获一个地理场所给人们带来的体验。

5. 空间交互分析

在地理学研究中，空间交互(spatial interaction)指的是两个场所之间的联系，通常可

以基于人流、货流、资金流等进行量化。研究空间交互有助于理解一个区域内部的结构以及动态演化特征。在空间大数据中，个体的移动轨迹以及个体之间的社交关系都可以在聚集层面量化两个场所之间的交互强度，前者如两个城市间的人流总量，后者如两个城市之间互相关注的好友对数。空间交互强度受到距离衰减效应的影响，距离远的两个地理单元间的联系相对较弱。因此，在地理学研究中，大多基于重力模型来拟合场所之间的交互强度，采用距离的负幂函数($d-\beta$)表示空间阻隔的影响。

目前可用的拟合方法有线性规划法、代数求解法、模拟法等。根据重力模型拟合结果，可以通过距离衰减系数 β 来表征特定空间交互行为中距离衰减效应的大小，即 β 值越低，距离的影响越小。实证研究表明，对于居民在城市尺度的移动行为，距离衰减系数在 1～2 之间，而对利用手机、社交媒体等途径建立的空间交互，距离衰减效应尽管较弱($\beta<1$)，但依然存在影响。

利用地理单元之间的空间交互，可以构建嵌入空间的网络（spatially-embedded network)，并引入网络分析方法研究其结构特征。在该网络中，通常每个结点为一个地理单元，而边的权重为地理单元间交互的强度，基于空间交互，构建嵌入空间的网络，从而引入网络科学分析方法，分析研究区的空间结构特征。在复杂网络研究中，常见的分析方法是对网络进行社区发现(community detection)分析，而网络中的社区由相对联系更为紧密的结点构成。对于嵌入空间的网络而言，一个社区往往对应地理空间中联系相对紧密的区域。由于距离衰减效应以及行政区划的影响，如果仅仅考虑交互强度而不考虑相邻约束，社区发现的结果通常为空间上连续的区块，并且往往与行政区划边界相一致。

城市是空间大数据产生最频繁的区域。因此，空间大数据的应用研究目前主要集中在城市区域。相关的研究领域有交通管理、城市规划、环境、公共卫生等。在此基础上，郑宇等提出了城市计算(urban computing)的概念，利用包括空间大数据在内的城市多源数据进行计算分析，发现并解决城市运行中的问题。

在上述应用中，除了空间大数据外，还要结合传统空间数据(如城市用地和建筑数据、道路网数据、检测站点数据等)进行分析。例如，有学者利用旧金山和波士顿地区的手机数据和路网数据，发现了交通拥堵路段的车流来源，并且给出了缓解拥堵的建议；有学者利用监测站数据、天气数据以及交通和人的移动数据，推断城市的实时精细分辨率空气质量数据，该结果有助于城市居民规划户外活动。由于空间大数据的获取建立在海量群体的空间行为的基础上，因此使我们能够更好地感知人的行为模式及其与地理环境之间的耦合模型。可以认为建立在社会感知基础上的公共政策制定，更能够体现“以人为本”的理念，有着广阔的应用前景。

空间大数据为我们提供了一条通过海量人群的空间行为模式去观察、理解地理环境特征及影响的研究路径。社会感知概念的提出正是概括了空间大数据的这种能力。空间大数据的处理，一方面需要有高效的分析方法，另一方面需要对人的行为动力学模型和地理环境特征有充分的理解。因此，需要信息科学、复杂性科学、地理学等不同学科以及不同应用领域的学者进行通力合作，才能有效提取空间大数据中所蕴含的信息，并充分体现其应用价值。

7.6.2 基于社会媒体的预测技术

社会媒体对预测的作用有两方面。一是社会信号的采集。例如，如果发现社会媒体上某一特定区域的人群都在发布信息说："我感冒了"，那么，这一区域很有可能正在传播流行性疾病，且有爆发的趋势。二是大众预测的融合。例如，美国大选期间，推特(Twitter)和脸谱(Facebook)在网上掀起预测热潮，很多网友在社会媒体上发布自己的预测结果，这种预测反映了社会媒体的群体智慧。

准确的预测结果对于人们在生活中的趋利避害、工作计划决策起着至关重要的作用。一个决策产生的结果与该决策本身有着时间上的滞后关系，"利"与"害"总是存在于未来的时间与空间中，任何决策都不可避免地要依赖于预测。对未来趋势提前做出判断，有利于适时地调整计划以及采取措施实施调控。

人类的预测活动分为自然预测和社会预测，分别面向自然界和人类社会。二者又存在较大差异，主要表现在主客体关系、规律性质、复杂程度和不确定性程度等几个方面，如表 7.1 所示。

表 7.1 自然预测与社会预测的区别

比较方面	自然预测	社会预测
主客体关系	自然的运行不因被预测而受干扰	互动反射关系(因应行为)，复杂博弈关系
规律性质	承认规律，了解事实	承认规律，了解事实
复杂程度	小	大
不确定性	小	受力面多，不确定性大
举例	天气变化、地震等	电影票房、总统大选等

自然预测的客体是自然现象，自然现象对人类的预测毫无感知能力，其运行轨迹不会因为预测而受到任何干扰。而社会预测的客体本身也是人，人会对预测结果产生因应行为。所谓因应行为，是指被预测的客体根据预测结果调整自己的行为，使得预测结果不准。相对而言，社会要比自然的"受力面"多得多，因而不确定性也大得多，对其进行预测也愈加困难。社会作为一个由大量子系统组成的非线性动态系统，在特定情况下会对某些微小的变量极为敏感。基于社会媒体的预测是指研究人类广泛参与并与社会发展变化有关的预测问题。

这种预测研究在许多领域都有着广泛的应用，例如金融市场的走势预测、产品的销售情况预测、政治大选结果预测、自然灾害的传播预测等。以往基于社会媒体的预测研究工作主要关注的是相关关系的发现和使用，通过找到一个现象的良好关联物来帮助了解现在和预测未来。例如，根据"微博声量"以及用户的情感分析可以预测股票的涨跌、电影票房的收入以及大选结果等。

我们需要站在一个全新的视角，介绍基于消费意图挖掘的预测以及基于事件抽取的预测，并通过挖掘影响预测客体未来走势的本质原因进一步提高预测精度。

在图 7.18 中，基于社会媒体的预测技术需要相关关系和因果关系的共同支撑，相关

关系可以从微博声量统计、情感倾向性分析、话题抽取等方面考虑,也可以运用更复杂的自然语言处理技术,从相关事件的抽取和消费意图的挖掘方面进行研究。因果关系对预测的帮助包括"由因导果"和"执果溯因"两方面,前者是正向地利用因果关系进行预测,后者是在预测失效时逆向找出失效的原因。

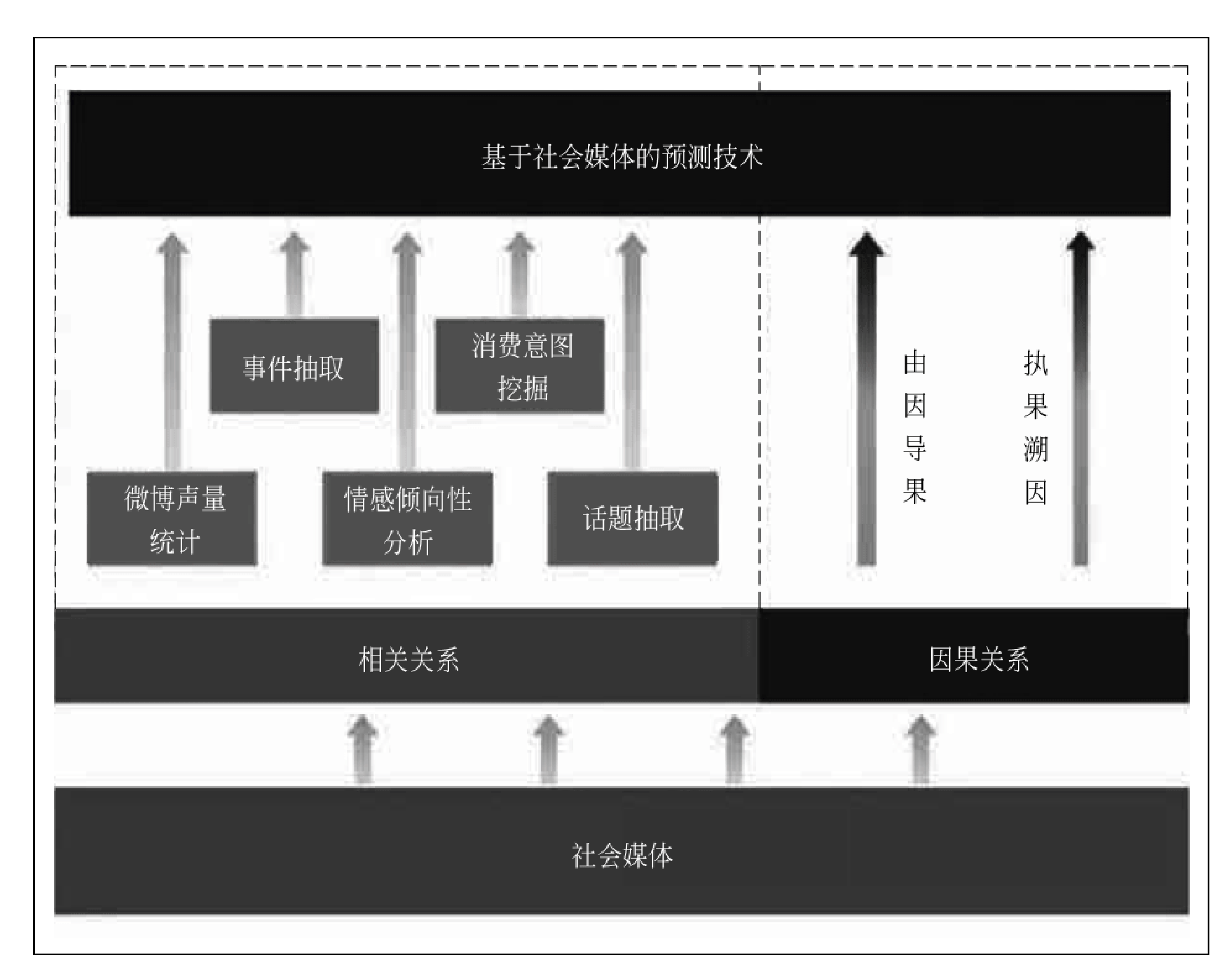


图 7.18　挖掘影响预测客体未来走势的本质原因

7.6.3　基于消费意图挖掘的预测

1. 基于社会媒体的消费意图挖掘

消费意图是指消费者通过显式或隐式的方式来表达对于某一产品或服务的购买意愿。社会媒体用户多,发布的信息量大。在这些信息中,用户会表达各种各样的需求和兴趣爱好。从大量的观测数据中,我们发现相当比例的社会媒体文本直接包含了用户的某种消费意图,例如:

"体感游戏还不错,考虑入手。"

"好想看《匆匆那年》啊!"

"我儿子 1 岁了,医生说有点缺钙,需要给孩子吃点什么呢?"

"天气转冷,换衣的季节到了,今年流行什么款式和颜色?"

第 1 条表达了用户想买体感游戏机,第 2 条表达了用户想去看电影《匆匆那年》,第 3 条要买补钙产品,第 4 条想买冬装。如果能够很好地挖掘出社会媒体用户对于某一产品的购买意愿,那么对于预测该产品的销量将有重要意义。

消费意图可分成"显式消费意图"和"隐式消费意图"两大类。显式消费意图是指在用

户所发布的微博文本中,显式地指出想要购买的商品,如第1、2两个例子。而隐式消费意图是指用户不会在所发布的微博文本当中显式地指出想要购买的商品,需要阅读者通过对文本语义的理解和进一步推理才能够猜测到用户想要购买的商品,如第3、4两个例子。

对于显式消费意图,很多学者通过模式匹配的方法识别。例如,在识别观影意图时,基于依存句法分析结果构建模板,识别对某部电影具有显式观影意图的微博,其准确率可以达到80%左右。而隐式消费意图的识别则难得多,难点包括:

(1) 如何理解用户的语义文本,进而理解用户的消费意图。这需要我们很好地理解和整合词汇级的语义特征以及句子级的语义特征。例如,要想识别出“我儿子1岁了,医生说有点缺钙,需要给孩子吃点什么呢”这句话包含的消费意图,需要理解关键词“儿子”、“缺钙”以及整个句子的含义。

(2) 用户消费意图的挖掘任务是领域相关的,因此构建的模型需要具有领域自适应能力。

为了解决以上难点,文献首次提出了基于领域自适应卷积神经网络的社会媒体用户消费意图挖掘方法。卷积神经网络对于解决该任务有以下两方面的优势:

(1) 卷积神经网络中的卷积层可以以滑动窗口的方式捕捉词汇级语义特征,而马克斯池(max pooling)层则可以很好地将词汇级特征整合成句子级语义特征;

(2) 卷积神经网络可以学习不同层次的特征表示,而一些特征表示则可以在不同领域间迁移。

消费意图毕竟还只是停留在个人意愿层面,有多少用户会真正将消费意图转化成消费行为,这是我们更加关心的话题,也是对于预测更有效的特征。消费意图识别的研究分成显式消费意图、隐式消费意图和能够转化成行为的意图三个层次。如图7.19所示,显式消费意图是用户消费意图这座冰山中露出水面的一角,大部分是隐式意图。而无论是显式意图,还是隐式意图,都只有一部分能够转化为购买行为。

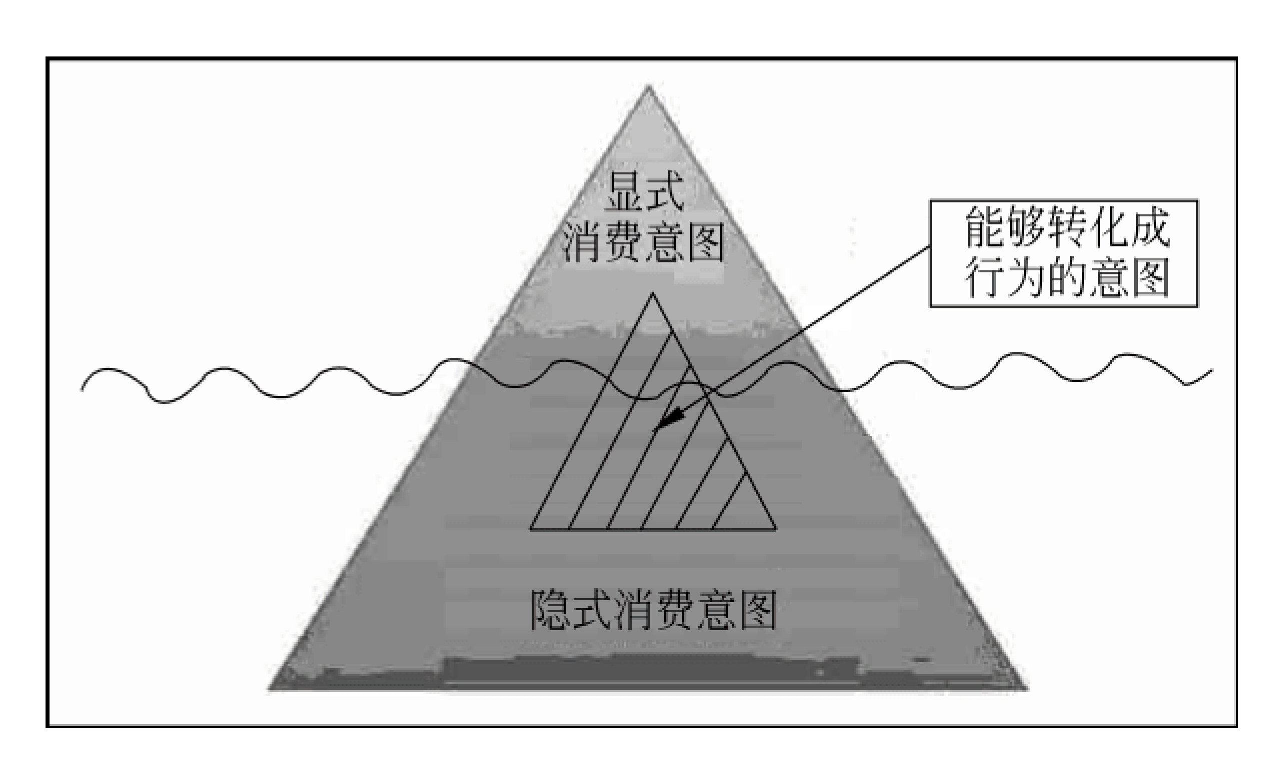


图7.19 消费意图研究层次

2. 基于消费意图挖掘的电影票房预测

消费意图挖掘在很多方面都有重要应用,如推荐系统、产品销量预测等。电影票房预测正是消费意图研究的一个成功应用:

很多与电影相关的数据可以方便地获取到。互联网上有很多与电影主题相关的网

站，例如美国电影资料库（Internet Movie Database，IMDB）、中国时光网、豆瓣网等。新浪微博每周至少会有 1000 万条以上的消息讨论与电影相关的内容。因此，有足够的数据用于分析影响电影票房的因素。

电影的总票房、周票房甚至是每天的票房都可以比较容易地从 IMDB 或网票网上获得，这有助于我们评价实验结果的好坏，并不断提高预测准确率。

社会媒体的消费意图数据与电影票房有清晰的逻辑相关性。社会媒体用户在某部电影上映前发布了关于某部电影的消息，说明他对这部电影感兴趣并且很有可能会去电影院观看这部电影。上映前一周的社会媒体数据相对于其他时间段的数据来讲，与电影票房的关联性最强。电影上映之后，带有情感倾向性的社会媒体内容变得至关重要。因为这类信息的传播可以看成是一种口碑营销，它将在很大程度上影响潜在消费者。

基于消费意图理解的电影票房预测相对于传统的电影票房预测而言，可以说是站在一个全新的角度进行研究。传统电影票房预测始于 20 世纪 80 年代末，美国电影经济学家巴瑞・利特曼（Barry Litman）在其论文《电影经济成功预测：基于 80 年代人的经验》（Predicting Financial Success of Motion Pictures：The 80's Experience）中首次提出了电影票房研究的基本模型和方法。总体来讲，传统电影票房预测主要是基于电影相关的特定的结构化数据，比如影片类型、美国电影协会分级、上映时间、是否有续集等。然而，这些方法要么预测效果不佳，要么需要一些时间点之后的数据才能得出合理的预测结果，很难被应用于实践中。

近几年，一些工作向人们展示了社会媒体在预测方面惊人的力量。例如，基于社会媒体的选举结果预测、流行病预测、奥斯卡获奖预测、足球比赛结果预测等。美国惠普实验室首先在基于社会媒体的电影票房研究中进行了尝试，在他们的研究中有两个重要的假设：一个是电影在社会媒体中被提及的次数（声量）越多，电影票房会越高；另一个是社会媒体用户对电影的评价越高，电影票房越高。但是，我们仔细分析后发现这两个假设并不成立。因为电影的媒体声量大并不一定意味着电影的口碑好；电影的口碑好，看的人不一定就多，口碑差，看的人不一定就少。真正能够做到口碑与票房双赢的电影并不多。

例如，《三枪拍案惊奇》《画皮》等电影的口碑较低（豆瓣评分 4.6 分），但是票房收入不错（票房收入分别是 2.6 亿元和 1.6 亿元）。我们认为，无论某个产品在社会媒体上被讨论得多么热烈，评价多么好，最终有多少人愿意购买才是影响产品销量最本质的因素。另外，对于像电影票房这样的预测对象，是需要在产品发布之前给出预测结果的。

然而，在产品发布之前没有产品的口碑数据，我们只能获得大众对该产品的消费意图数据（购买意愿）。因此，基于消费意图的电影票房预测打破了以往的格局限制，从最根本的因素出发来预测电影票房收入。

电影票房预测的主流模型可分为线性预测模型和非线性预测模型。这两个模型都存在一个前提，即认为电影票房收入与预测影响因素之间存在线性或非线性关系。在首周票房预测实验中，线性回归模型实验结果要好于非线性回归模型，而在总票房预测研究中，非线性回归模型效果要优于线性回归模型。这表明电影上映前一周的数据与首周票房线性关系比较明显，这时线性回归模型的预测能力要高于非线性回归模型。随着时间的推移，各种新的因素不断加入以及一些偶然情况的发生，使得电影上映前一周的数据与

总票房之间的线性关系越来越不明显，而这时线性回归模型的预测能力就要低于非线性回归模型。将线性回归模型和非线性回归模型相结合是相关研究未来的一项重要工作。

7.6.4 基于事件抽取的预测

基于消费意图的预测是从人的主观角度出发进行预测，而基于事件的预测则是从客观的事实角度出发进行预测。社会媒体中报道的一些事件会对人们的决策产生影响，而人们的决策又会影响到他们的交易行为，这种交易行为最终会导致金融市场的波动。重要事件会导致股票市场的剧烈震荡，如果能够及时准确地获取这些重要事件，势必会有助于对金融市场波动的预测。

金融市场的预测研究可分成时间序列交易数据驱动和文本驱动两个不同方向。

时间序列交易数据是最早用于建立预测模型的一类数据，主要包括股票历史价格数据、历史交易量数据、历史涨跌数据等。在传统的金融市场预测研究中，金融领域学者多从计量经济学的角度出发进行时间序列分析，进而预测市场的波动情况。

文本驱动的金融市场预测主要是挖掘新闻报道和社会媒体中报道的客观事实以及大众的情感波动。前人的很多研究工作表明，金融领域的新闻在一定程度上会影响股票价格的波动。之后自然语言处理技术逐渐被引入到金融市场预测中。而早期被应用在文本表示的技术主要是基于词袋模型(bag-of-words)。有文献指出，基于词袋模型的文本表示方法并不是最优方案，基于语义框架可以挖掘出更加丰富的文本特征。

以上工作存在一个共性的问题，即没有提取文本中的结构化信息，而这一信息对于股票涨跌预测非常重要。例如，“甲骨文公司诉讼 Google 公司侵权”，如果用词袋模型表示，其形式为{“甲骨文”,“诉讼”,“Google”,“侵权”,…}。我们从中并不能判断出是甲骨文公司诉讼 Google 公司，还是 Google 公司诉讼甲骨文公司，也就很难判断出哪个公司的股价会上涨或下跌。

有一种想法是利用结构化的事件预测股票的涨跌。对于上面的例子，如果利用结构化的事件，则可以表示成{(施事：“甲骨文”),(行为：“诉讼”),(受事：“Google”)}。由此，我们能够清楚地知道是甲骨文公司诉讼 Google 公司。在此基础上可预测 Google 公司的股价有可能受影响而下跌，而甲骨文公司的股价可能会上涨。

7.6.5 基于因果分析的预测

对于许多预测问题来说，因果分析是十分重要并且高效的。与相关性相比，因果的确定性更强。例如疾病预测、行为预测和政策效用预测等。对于某些事件来说，当没有过多的相关性数据可用时，因果是最有效的预测指南。例如稀有事件预测、新闻事件预测等。当基于相关性的预测失效时，因果更是预测的唯一指南。因此，当我们对于某一事物预测不准或者认识不准时，一个合理的做法是分析因果并使用因果进行再认识。

1. 因果关系概述

原因与结果是重要的哲学范畴。对事物间因果关系的探索，自人类诞生以来就开始了。因果关系也是人类在漫长的社会实践中逐步总结出来的一个基本法则，成为人们推

理事实和认识未知的指南。以下把因果视为关系、知识和逻辑。

1）因果是关系

作为一种语义关系，因果关系是语义理解和篇章分析的重要资源。

2）因果是知识

因果作为一种重要的知识形式，是问答系统和决策的重要依据和资源。要回答“是什么导致肿瘤缩小”这类问题，一个大型的因果关系知识库是必要的。对于一个现象或者状况的出现，只有知道导致它出现的原因，才能根据原因提出相应的对策。作为决策依据的因果是区别于相关的本质特性。

3）因果是逻辑

作为逻辑的因果，是因果最重要的方面。作为科学逻辑中最重要的组成部分，因果逻辑体现在预测逻辑和解释逻辑两个方面。

因果与相关是两个不同的重要概念，尽管在很多科学研究中因果比相关更重要，但是目前大数据侧重于相关性研究。相关性分析得到的结论有时是不可靠的，甚至是错误的。无因果关系的两个变量之间可能会表现出虚假的相关性。很多例子可以说明虚假相关性，如张三和李四的手表上的时间具有很强的相关性，但是人为地改变张三的手表时间，不会引起李四的手表时间的变化。

统计上的研究表明，小学生的阅读能力与鞋的尺寸有很强的相关性，但是很明显它们没有因果关系，人为地改变鞋的尺寸，不会提高小学生的阅读能力。

因果关系也可能表现出虚假的独立性。统计表明：练太极拳的人平均寿命等于或者低于不练太极拳的人。事实上，太极拳确实可以强身健体、延长寿命，但练太极拳的人往往是体弱多病的人，所以表现出虚假的独立性。

因此，表面上相关的事情，实质上可能并无关联，更没有因果的必然性；表面上不相关，但可能背后有因果关系。大数据分析不能只考虑相关性，也应该考虑因果关系。

如图 7.20 所示，A 代表“气温”，B 代表“冰激凌销量”，C 代表“游泳馆客流量”。A 是 B 和 C 的共同原因，A 升高会导致 B 和 C 的增加。虽然 B 与 C 存在统计相关性，但如果想提高 B 显然不能通过干预 C 来达到，而能通过 A 的升高来达到。

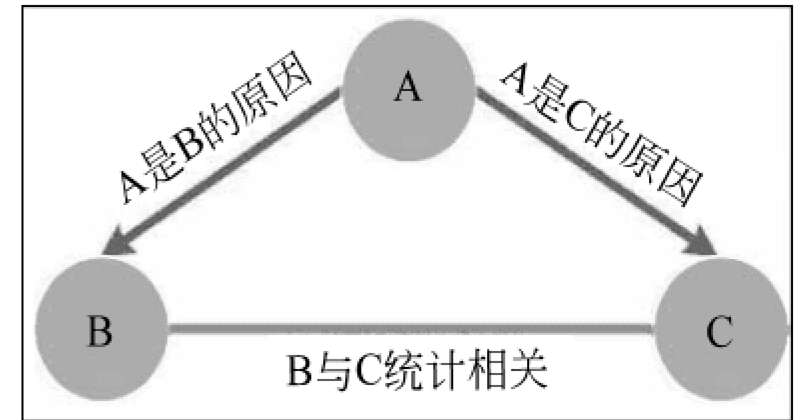


图 7.20　因果关系与相关关系的区别

2. 因果关系抽取

因果关系抽取是一个非常基础且重要的工作。抽取出的因果关系或因果知识可用于预测、问答等。在文本中进行因果抽取就要用到自然语言的处理技术和方法，如词性标注、句法分析、短语抽取等。对于因果关系抽取和检测任务来说，前人的工作所使用的线索可以粗略地分为三类：

1）上下文词信息

在自然语言文本中，相同或相似的句法结构对应不同的语义关系，上下文信息对区别这种相同或相似句法结构的不同语义关系具有重要意义。文献[20]指出，丰富的上下文

信息对提高因果抽取的准确率是非常必要的。获得含有因果提及的句子,尤其是含有显式因果提及的句子是相对容易的。

2)词之间的关联信息

虽然使用因果关系触发词能覆盖大多数情况,但如果从含有因果提及的句子中抽取出真正存在因果关系的“词对”或者“事件对”是比较困难的。有文献认为因果提及中的名词之间、动词之间、动词和名词之间的关联信息对于识别因果来说是非常有效的资源。因此提出了一种基于分布式相似性的半指导因果事件的识别算法。

3)动词和名词的语义关系信息

在自然语言中一些词语本身蕴含着因果关系的可能性,例如英文的 Increase X、Decrease X、Cause X、Preserve X 都很可能激发出一个原因的结果;中文的“增加了×”“避免了×”“防止了×”也具有同样的功能。这些词一般被称为触发词。

基于这种触发词模板方法进行因果关系抽取的工作有很多。例如文献[23],通过把这些作为谓语动词的触发词模板人工地分为 CAUSATION、MATERIAL、NECESSITY、USE、PREVENTION 五类,来区分抽取到的因果关系的类型;文献[24]使用因果关系触发词抽取文本中的名词因果对,使用这种因果对来判断一个句子是否是描述因果逻辑的句子;一些文献则利用因果关系词在大量的新闻语料中获取事件之间的因果关系。

3. 由因导果

“由因导果”即因果的预测逻辑。看到一个现象或者一个事件的发生,我们总想知道未来可能出现的现象或者发生的事件。对于预测未来,因果无疑是最有效的指南和依据。尤其是在基于相关性分析的预测失效时,若能分析出原因并利用原因进行预测,则预测结果会更加可靠。

通过抽取大规模新闻语料中新闻事件和事件之间的因果关系,有文献把这些因果事件分类、关联,并组成事件因果关系网络,使用这个网络预测未来事件。所有的因果事件都表示成因果“事件对”的形式,其中原因事件和结果事件都尽量用六元组形式表示。通过计算因果“事件对”之间的相似性来预测结果事件。

在利用因果来做预测的工作中,事件通常采用的是名词短语或 n 元组的表示形式。但基于这种表示形式来做事件的匹配,会漏掉很多事件本身的信息,从而导致匹配的效果不好。另一类问题是稀有事件的预测。稀有事件是指发生概率很低的事件。例如,公路交通事故、网络欺诈行为、网络入侵行为、信用卡诈骗行为等。稀有事件的预测是一个非常复杂的问题,它需要有对问题本身的深刻理解和对问题中的不确定性进行建模。对于预测稀有事件,数据的稀疏性导致缺少大量的相关关系或相关事件。因此,对稀有事件的预测,既需要具备正确的因果知识,又要能够进行正确的因果分析,同时还能充分利用可以用到的小样本数据。

4. 执果溯因

“执果溯因”即因果的解释逻辑。看到一个现象或一个结果时,我们总想知道“为什么”。在自然语言文本中,我们对因果解释逻辑的诉求也是随处可见。以电商为例,电商网站上有大量用户对商品的评论信息,如某些人对商品 A 持有积极评价,另一些人则对

商品 A 吐槽。作为生产商和销售商很想知道，为什么有些人喜欢，而有些人不喜欢。如果能从评论数据中进行分析找到原因，对生产商和销售商来讲都有重大意义。

在社会学和大众舆情分析领域，大众对某个社会事件或者社会问题的情感和态度是十分重要的，但是更重要的是大众持有某种情感或者态度的原因。如果能自动地从文本中尤其是社会媒体文本中挖掘出这些原因，这对于理解民意、维护社会安定具有重大意义。类似这种从文本中分析原因的需求几乎覆盖各行各业。

在商业决策领域，我们想知道产品销量提高或者降低的原因，进而做出应对，例如电影票房的涨跌和广告宣传的因果作用分析对于宣传策略的选择至关重要。在政治决策上同样如此。为了分析一个时序变量是否对另一个时序变量产生因果作用，有文献提出了一个基于贝叶斯网络的时间序列模型。先预测出一个虚拟结果，进而和真实结果进行对比来评价一个变量对另一个变量的因果作用。比如有一个网站，在某一时刻 t 加入了一个广告，那么这个广告究竟可以带来了多少点击量？

如图 7.21 所示，竖切的虚线代表引入广告的分界线，original 部分的实线和虚线分别表示真实的网站点击量曲线和不引入广告的情况下的网站点击量曲线（反事实点击量曲线，通过预测得到）。pointwise 部分代表的是真实曲线和反事实曲线的差值曲线。cumulative 部分是真实曲线和反事实曲线累积差值。通过观察累积差值的大小，可以得

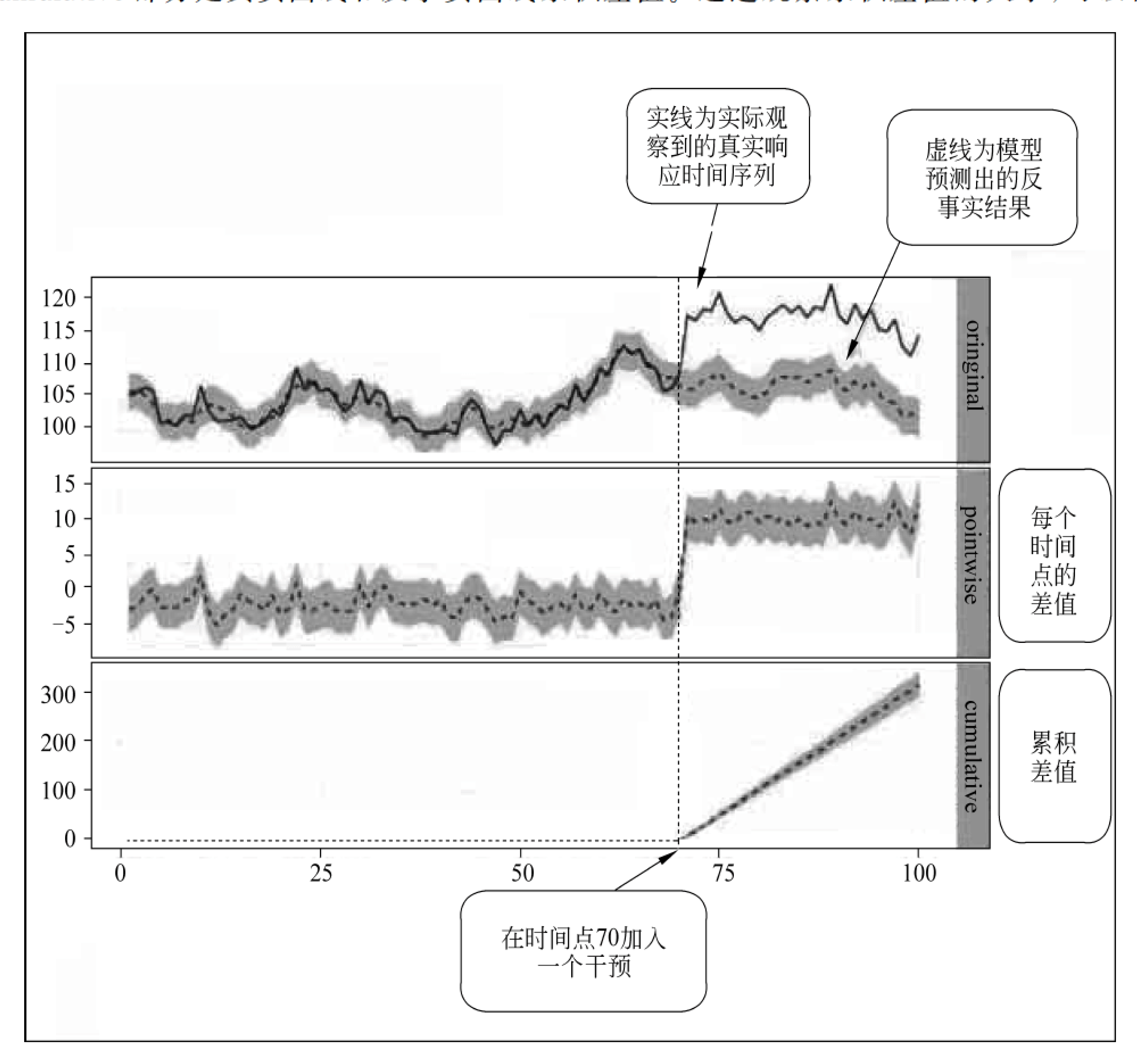


图 7.21　通过反事实结果预测推断因果效用

到引入广告对网站点击量增加的因果效用，比如得出“引入广告是网站点击量显著增加的原因”的结论。

7.7 大数据应用案例之：如何用大数据看风水？以星巴克和海底捞的选址为例

有人问：你们整天说大数据，它到底有啥用啊？下面先介绍一下如何用大数据来看“风水”！

说起看风水开店选址，大家脑海里浮现出来的十有八九是风水先生们拿着罗盘走来走去的画面。

而在互联网时代，商家们紧跟时代步伐已经学会了用大数据看“风水”。简单说来就是基于搜索数据来推断出来哪个地方的用户对服务和商品有需求，相当于是根据需求的密集程度来选址——这大概是开店选址最关键的一步，也是百度大数据最独特的地方。

举个例子，下面是一份研究的是星巴克和海底捞未覆盖地区的用户对这两家店的需求分析的数据图表如图 7.22 和图 7.23 所示。



图 7.22 星巴克和海底捞

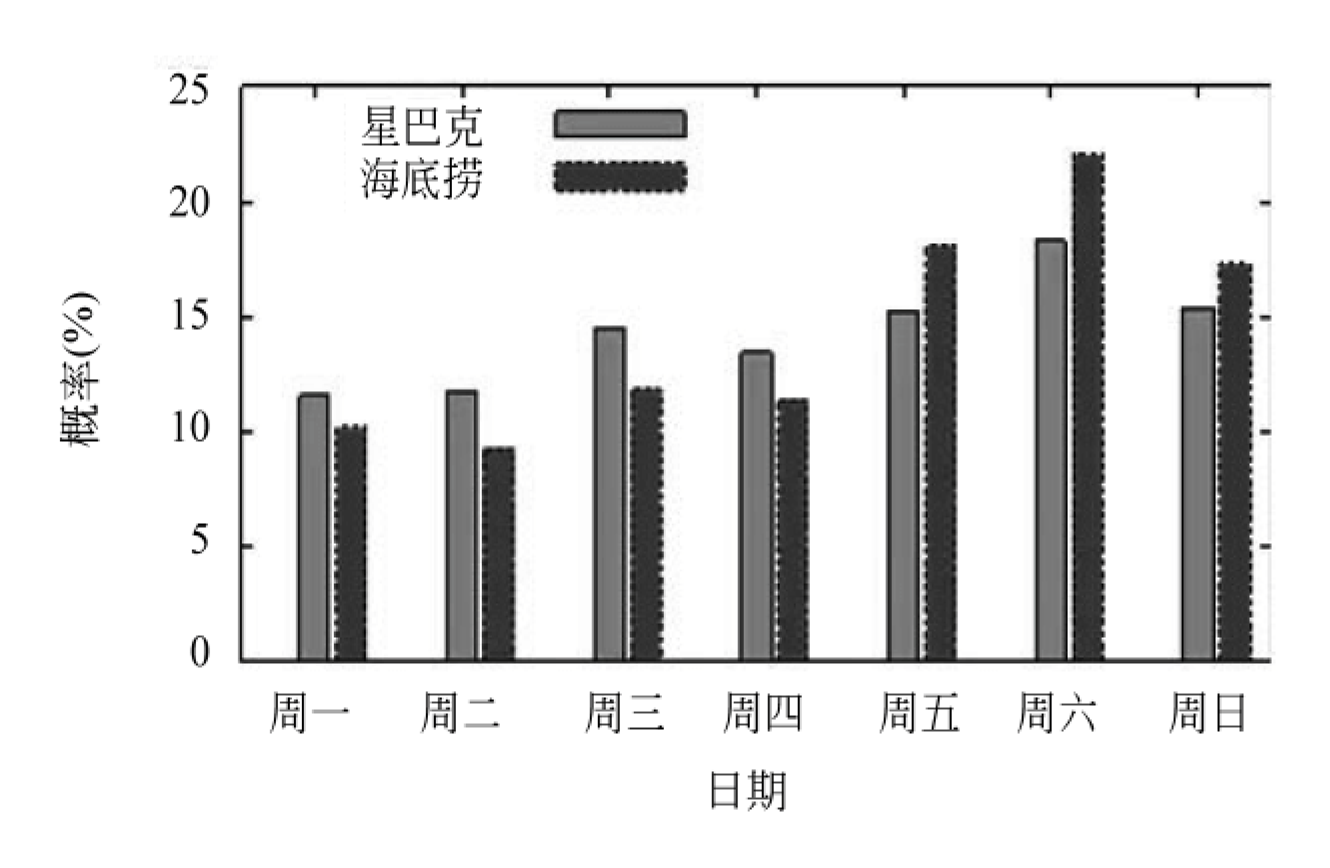


图 7.23 星巴克和海底捞未覆盖地区的用户对这两家店的需求分析

看完之后是不是发现看不懂？没关系，我们已经为你翻译好了：

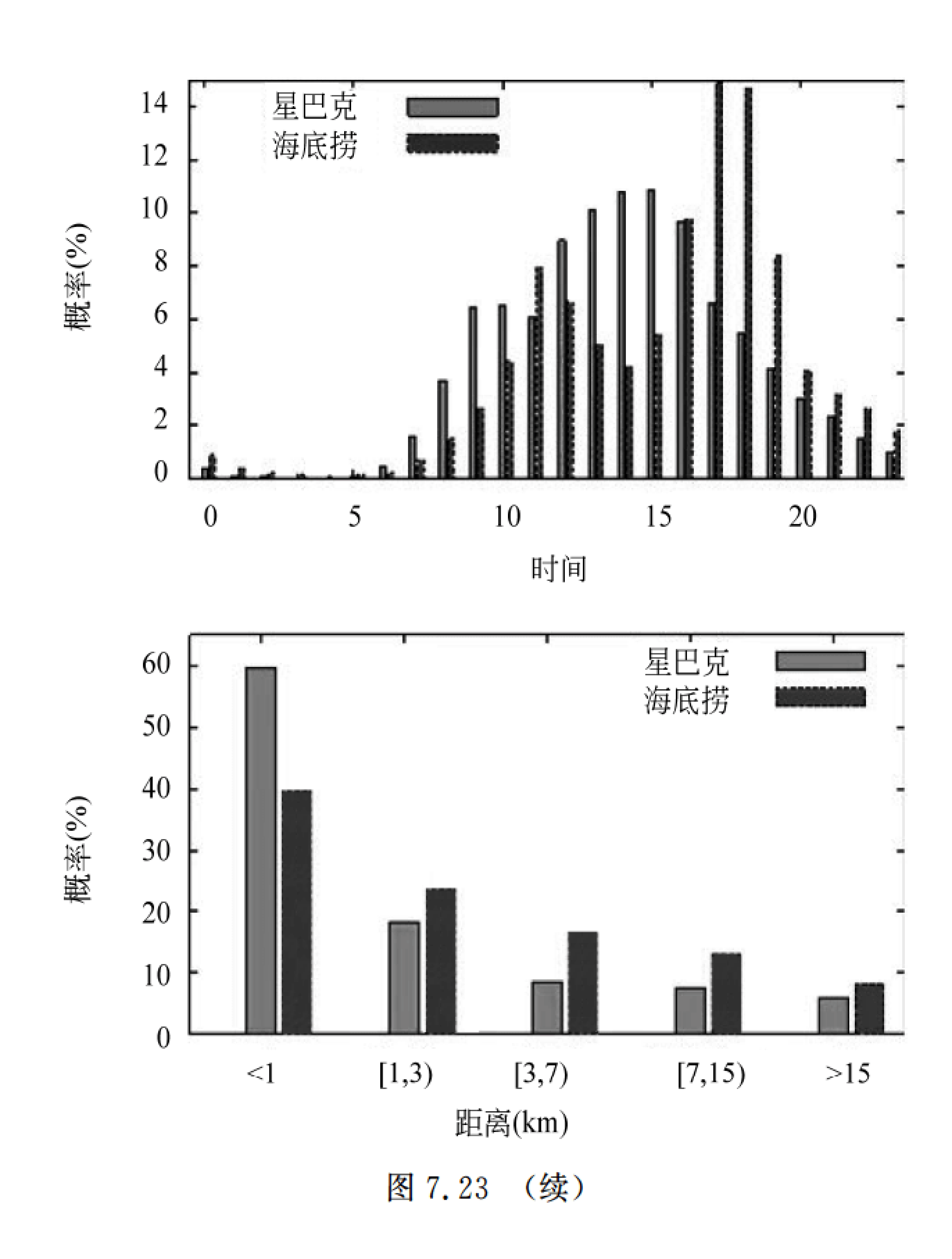


图 7.23　(续)

图一：对比一周的需求，吃货们在周末对海底捞的需求高过星巴克。

图二：在一天之内，单身狗喜欢在午饭后约女神喝星巴克。

图三：七成星巴克消费者一般选在附近 1km，而吃海底捞一般需要跑更远的距离(大约 3 公里)。

Big Data Lab(百度大数据实验室)要做的就是通过分析这些时间、空间、网点、交通便利程度、竞争对手情况等等因素，结合用户需求，告诉你应该在哪里开店。

习题与思考题

一、选择题

1. 某超市研究销售记录数据后发现，买啤酒的人很大概率也会购买尿布，这种属于数据挖掘的哪类问题？(　　)

A. 关联规则发现　　B. 聚类　　C. 分类　　D. 自然语言处理

2. 数据挖掘的挖掘方法包括(　　)。

A. 聚类分析　　B. 回归分析　　C. 神经网络　　D. 决策树算法

3. Web 内容挖掘实现技术(　　)。

A. 文本总结　　B. 文本分类　　C. 文本聚类　　D. 关联规则

4. 社交网络产生了海量用户以及实时和完整的数据，同时社交网络也记录了用户群

体的(　　),通过深入挖掘这些数据来了解用户,然后将这些分析后的数据信息推给需要的品牌商家或是微博营销公司。

A. 地址　　B. 行为　　C. 情绪　　D. 来源

5. 文本挖掘的工具有(　　)。

A. SPP Text Mining　　B. IBM DB2 intelligent Miner

C. SAS text miner　　D. SPSS Text Mining

6. 数据挖掘工作的四个阶段,数据挖掘占总时间的百分比(　　)%,对于成功重要性的百分比(　　)%。

A. 50　　B. 20　　C. 80　　D. 60

7. 美国海军军官莫里通过对前人航海日志的分析,绘制了新的航海路线图,标明了大风与洋流可能发生的地点。这体现了大数据分析理念中的(　　)。

A. 在数据基础上倾向于全体数据而不是抽样数据

B. 在分析方法上更注重相关分析而不是因果分析

C. 在分析效果上更追究效率而不是绝对精确

D. 在数据规模上强调相对数据而不是绝对数据

8. 下列关于舍恩伯格对大数据特点的说法中,错误的是(　　)。

A. 数据规模大　　B. 数据类型多样

C. 数据处理速度快　　D. 数据价值密度高

9. 下列关于聚类挖掘技术的说法中,错误的是(　　)。

A. 不预先设定数据归类类目,完全根据数据本身性质将数据聚合成不同类别

B. 要求同类数据的内容相似度尽可能小

C. 要求不同类数据的内容相似度尽可能小

D. 与分类挖掘技术相似的是,都是要对数据进行分类处理

10. 下列关于大数据的分析理念的说法中,错误的是(　　)。

A. 在数据基础上倾向于全体数据而不是抽样数据

B. 在分析方法上更注重相关分析而不是因果分析

C. 在分析效果上更追究效率而不是绝对精确

D. 在数据规模上强调相对数据而不是绝对数据

11. 建立在相关关系分析法基础上的预测是大数据的(　　)。

A. 基础　　B. 前提　　C. 核心　　D. 条件

12. 关于数据创新,下列说法正确的是(　　)。

A. 多个数据集的总和价值等于单个数据集价值相加

B. 由于数据的再利用,数据应该永久保存下去

C. 相同数据多次用于相同或类似用途,其有效性会降低

D. 数据只有开放价值才能得到真正释放

13. 关于数据估值,下列说法错误的是(　　)。

A. 随着数据价值被重视,公司所持有和使用的数据也渐渐纳入了无形资产的范畴

B. 无论是向公众开放还是将其锁在公司的保险库中，数据都是有价值的
C. 数据的价值可以通过授权的第三方使用来实现
D. 目前可以通过数据估值模型来准确地评估数据的价值评估

14. 以下哪种说法是错误的(　　)。
A. 将罪犯的定罪权放在数据手中，借以表达对数据和分析结果的崇尚，这实际上是一种滥用
B. 随着数据量和种类的增多，大数据促进了数据内容的交叉检验，匿名化的数据不会威胁到任何人的隐私
C. 采集个人数据的工具就隐藏在我们日常生活所必备的工具当中，比如网页和智能手机应用程序
D. 预测与惩罚，不是因为所做，而是因为将做

15. 对大数据使用进行正规评测及正确引导，可以为数据使用者带来什么切实的好处？(　　)。
A. 他们无须再取得个人的明确同意，就可以对个人数据进行二次利用
B. 数据使用者不需要为敷衍了事的评测和不达标准的保护措施承担法律责任
C. 数据使用者的责任不需要强制力规范就能确保履行到位
D. 所有项目，管理者必须设立规章，规定数据使用者应如何评估风险、如何规避或减轻潜在伤害

二、问答题

1. 大数据分析面对的数据类型有哪些？
2. 简述大数据分析与处理方法。
3. 数据挖掘的功能有哪些？
4. 为什么说“大数据自动挖掘”才是大数据的真正意义？
5. 为什么说“商务智能＝数据＋分析＋决策＋利益”？
6. 电商大数据分析需要考虑哪些方面？
7. 简述大数据营销的定义与特点。
8. 谈一谈你对网络营销大数据业务模型和实际操作的看法。
9. 基于社会媒体的分析预测技术有哪些？

第 8 章　大数据隐私与安全

目前影响大数据产业发展主要有两个大问题：一个是大数据应用场景，一个是大数据隐私保护问题。

大数据商业价值的应用场景，大数据公司和企业正在寻找，目前在移动互联网的精准营销和获客、360 度用户画像、房地产开发和规划、互联网金融的风险管理、金融行业的供应链金融、个人征信等方面已经取得了进步，拥有了很多经典案例。

但在有关大数据隐私保护以及大数据应用过程中个人信息保护方面还停滞不前，大家都在摸石头过河，不知道哪些事情可以做，哪些事情不可以做。国家在大数据隐私保护方面正在进行立法，估计不久的将来，大数据服务公司和企业将会了解大数据隐私保护方面的具体要求。在没有明确有关大数据隐私保护法规前，我们可以参考国外的隐私法，严格遵守国际上通用的个人隐私保护法，在实施大数据价值变现的过程中，充分保护所有相关方的个人利益。

最后纵观人类历史，在任何领域，如果我们可以拿到数据进行分析，我们就会取得进步。如果我们拿不到数据，无法进行分析，我们注定要落后。过去因数据不足导致的错误远远好过那些根本不用数据的错误，因此我们需要掌握大数据这个武器，利用好它，帮助人类社会加速进化，帮助企业实现大数据的价值变现。

8.1　大数据面临的问题

大数据因为它所蕴含的潜在价值，正在成为企业的隐形“金矿”。随着生产、运营、管理、监控、销售、客服等各个环节的数据不断累积和增长，以及用户数的不断上升，通过从庞大的数据中分析出相关模式以及趋势，可以实现高效管理、精准营销，成为企业打开这一“金矿”的钥匙。然而传统的 IT 基础架构和数据管理分析方法已经不能适应大数据的快速增长。大数据的爆发是我们在信息化和社会发展中遇到的棘手问题，需要采用新的数据管理模式，研究和发展新一代的信息技术才能解决。大数据问题可归纳为表 8.1 中所列的 7 类。

8.1.1　大数据面临的安全问题

1. 速度方面的问题

传统的关系型数据库管理系统（RDBMS）一般都是集中式的存储和处理，没有采用分布式架构，在很多大型企业中的配置往往都是基于 IOE（IBM 服务器、Oracle 数据库、EMC 存储）。在这种典型配置中单台服务器的配置通常都很高，可以多达几十个 CPU

表 8.1　大数据问题

大数据问题分类	大数据问题描述
速度方面的问题	导入导出问题 统计分析问题 检索查询问题 实时响应问题
种类及架构问题	多源问题 异构问题 原系统的底层架构问题
体量及灵活性问题	线性扩展问题 动态调度问题
成本问题	大机与小型服务器的成本对比 原有系统改造的成本把控
价值挖掘问题	数据分析与挖掘问题 数据挖掘后的实际增效问题
存储及安全问题	结构与非结构 数据安全 隐私安全
互联互通与数据共享问题	数据标准与接口 共享协议 访问权限

核，内存也能达到上百 GB；数据库的存储放在高速大容量的磁阵上，存储空间可达 TB 级。这种配置对于传统的信息管理系统（MIS）需求来说是可以满足需求的，然而面对不断增长的数据量和动态数据使用场景，这种集中式的处理方式就日益成为瓶颈，尤其是在速度响应方面捉襟见肘。

在面对大数据量的导入导出、统计分析、检索查询方面，由于依赖于集中式的数据存储和索引，性能随着数据量的增长而急速下降，对于需要实时响应的统计及查询场景更是无能为力。比如在物联网中，传感器的数据可以多达几十亿条，对这些数据需要进行实时入库、查询及分析，传统的 RDBMS 就不再适合应用需求。

2. 种类及架构问题

RDMBS 对于结构化的、固定模式的数据，已经形成了相当成熟的存储、查询、统计处理方式。随着物联网、互联网以及移动通信网络的飞速发展，数据的格式及种类在不断变化和发展。在智能交通领域，所涉及的数据可能包含文本、日志、图片、视频、矢量地图等来自不同数据采集监控源的、不同种类的数据。

这些数据的格式通常都不是固定的，如果采用结构化的存储模式将很难应对不断变化的需求。因此对于这些种类各异的多源异构数据，需要采用不同的数据和存储处理模式，结合结构化和非结构化数据存储。在整体的数据管理模式和架构上，也需要采用新型的分布式文件系统及分布式 NoSQL 数据库架构，才能适应大数据量及变化的结构。

3. 体量及灵活性问题

如前所述,大数据由于总体的体量巨大,采用集中式的存储,在速度、响应方面都存在问题。当数据量越来越大,并发读写量也越来越大时,集中式的文件系统或单数据库操作将成为致命的性能瓶颈,毕竟单台机器的承受压力是有限的。我们可以采用线性扩展的架构和方式,把数据的压力分散到很多台机器上,直到可以承受,这样就可以根据数据量和并发量来动态增加和减少文件或数据库服务器,实现线性扩展。

在数据的存储方面,需要采用分布式可扩展的架构,比如大家所熟知的 Hadoop 文件系统和 HBase 数据库。同时在数据的处理方面,也需要采用分布式的架构,把数据处理任务分配到很多计算结点上,同时还需考虑数据存放结点和计算结点之间的位置相关性。在计算领域中,资源分配、任务的分配实际上是一个任务调度问题。其主要任务是根据当前集群中各个结点上面的资源(包括 CPU、内存、存储空间和网络资源等)的占用情况和各个用户作业服务质量要求,在资源和作业或者任务之间做出最优的匹配。由于用户对作业服务质量的要求是多样化的,同时资源的状态也在不断变化,因此,为分布式数据处理找到合适的资源是一个动态调度问题。

4. 成本问题

集中式的数据存储和处理,在对硬件软件选型时,基本采用的方式都是配置相当高的大型机或小型机服务器,以及访问速度快、保障性高的磁盘阵列,来保障数据处理性能。这些硬件设备都非常昂贵,动辄高达数百万元,同时软件也经常是国外大厂商如 Oracle、IBM、SAP、微软等的产品,对于服务器及数据库的维护也需要专业技术人员,投入及运维成本很高。在面对海量数据处理的挑战时,这些厂商也推出了形似庞然大物的“一体机”解决方案,如 Oracle 的 Exadata、SAP 的 Hana 等,通过把多服务器、大规模内存、闪存、高速网络等硬件进行堆叠,来缓解数据压力,然而在硬件成本上,更是大幅跳高,一般的企业很难承受。

新型的分布式存储架构、分布式数据库如 HDFS、HBase、Cassandra、MongoDB 等由于大多采用去中心化的、海量并行处理 MPP 架构,在数据处理上不存在集中处理和汇总的瓶颈,同时具备线性扩展能力,能有效地应对大数据的存储和处理问题。在软件架构上,也都实现了一些自管理、自恢复的机制,以面对大规模结点中容易出现的偶发故障,保障系统整体的健壮性,因此对每个结点的硬件配置,要求并不高,甚至可以使用普通的 PC 作为服务器,因此在服务器成本上可以大大节省,在软件方面开源软件也占据非常大的价格优势。

当然,在谈及成本问题时,我们不能简单地进行硬件软件的成本对比。要把原有的系统及应用迁移到新的分布式架构上,从底层平台到上层应用都需要做很大的调整。尤其是在数据库模式以及应用编程接口方面,新型的 NoSQL 数据库与原来的 RDBMS 存在较大的差别,企业需要评估迁移及开发成本、周期及风险。除此之外,还需考虑服务、培训、运维方面的成本。但在总体趋势上,随着这些新型数据架构及产品的逐渐成熟与完善,以及一些商业运营公司基于开源基础为企业提供专业的数据库开发及咨询服务,新型的分布式、可扩展数据库模式必将在大数据浪潮中胜出,从成本到性能方面完胜传统的集

中式大机模式。

5. 价值挖掘问题

大数据由于体量巨大，同时又在不断增长，因此单位数据的价值密度在不断降低。但同时大数据的整体价值在不断提高，大数据被类比为石油和黄金，因此从中可以发掘巨大的商业价值。要从海量数据中找到潜藏的模式，需要进行深度的数据挖掘和分析。大数据挖掘与传统的数据挖掘模式也存在较大的区别：

传统的数据挖掘一般数据量较小，算法相对复杂，收敛速度慢。然而大数据的数据量巨大，在对数据的存储、清洗、ETL（抽取、转换、加载）方面都需要能够应对大数据量的需求和挑战，在很大程度上需要采用分布式并行处理的方式，比如 Google、微软的搜索引擎，在对用户的搜索日志进行归档存储时，就需要多达几百台甚至上千台服务器同步工作，才能应付全球上亿用户的搜索行为。

同时，在对数据进行挖掘时，也需要改造传统数据挖掘算法以及底层处理架构，同样采用并行处理的方式才能对海量数据进行快速计算分析。Apache 的 Mahout 项目就提供了一系列数据挖掘算法的并行实现。在很多应用场景中，甚至需要挖掘的结果能够实时反馈回来，这对系统提出了很大的挑战，因为数据挖掘算法通常需要较长的时间，尤其是在大数据量的情况下，可能需要结合大批量的离线处理和实时计算才可能满足需求。

数据挖掘的实际增效也是我们在进行大数据价值挖掘之前需要仔细评估的问题。并不见得所有的数据挖掘计划都能得到理想的结果。首先需要保障数据本身的真实性和全面性，如果所采集的信息本身噪音较大，或者一些关键性的数据没有被包含进来，那么所挖掘出来的价值规律也就大打折扣。

其次也要考虑价值挖掘的成本和收益，如果对挖掘项目投入的人力物力、硬件软件平台耗资巨大，项目周期也较长，而挖掘出来的信息对于企业生产决策、成本效益等方面的贡献不大，那么片面地相信和依赖数据挖掘的威力，也是不切实际和得不偿失的。

6. 存储及安全问题

在大数据的存储及安全保障方面，大数据由于存在格式多变、体量巨大的特点，也带来了很多挑战。针对结构化数据，关系型数据库管理系统 RDBMS 经过几十年的发展，已经形成了一套完善的存储、访问、安全与备份控制体系。由于大数据的巨大体量，也对传统 RDBMS 造成了冲击，如前所述，集中式的数据存储和处理也在转向分布式并行处理。大数据更多的时候是非结构化数据，因此也衍生了许多分布式文件存储系统，分布式 NoSQL 数据库等来应对这类数据。

然而这些新兴系统，在用户管理、数据访问权限、备份机制、安全控制等各方面还需进一步完善。对于安全问题，简言之，一是要保障数据不丢失，对海量的结构、非结构化数据，需要有合理的备份冗余机制，在任何情况下数据不能丢；二是要保障数据不被非法访问和窃取，只有对数据有访问权限的用户，才能看到数据，拿到数据。

由于大量的非结构化数据可能需要不同的存储和访问机制，因此要形成对多源、多类型数据的统一安全访问控制机制，还是亟待解决的问题。大数据由于将更多更敏感的数

据汇集在一起，对潜在攻击者的吸引力更大；若攻击者成功实施一次攻击，将能得到更多的信息，“性价比”更高，这些都使得大数据更易成为被攻击的目标。LinkedIn 在 2012 年被曝 650 万用户账户密码泄露；雅虎遭到网络攻击，致使 45 万用户 ID 泄露。2011 年 12 月，CSDN 的安全系统遭到黑客攻击，600 万用户的登录名、密码及邮箱遭到泄露。

与大数据紧密相关的还有隐私问题。由于物联网技术和互联网技术的飞速发展，与我们工作生活相关各类信息都被采集和存储下来，我们随时暴露在“第三只眼”下面。不管我们是在上网、打电话、发微博、微信，还是在购物、旅游，我们的行为都在随时被监控分析。对用户行为的深入分析和建模，可以更好地服务用户，实施精准营销，然而如果信息泄露或被滥用，则会直接侵犯到用户的隐私，对用户形成恶劣的影响，甚至带来生命财产的损失。

2006 年，美国 DVD 租赁商 Netflix 公司举办了一个算法竞赛。该公司公布了大约来自 50 万用户的一亿条租赁记录，并且公开悬赏 100 万美元，举办一个软件设计大赛来提高他们的电影推荐系统的准确度，胜利的条件是把准确度提高 10%。尽管该公司对数据进行了精心的匿名化处理，还是被一个用户认出来了，一个化名“无名氏”的未出柜的同性恋母亲起诉了 Netflix 公司，她来自保守的美国中西部。

在美国的微博网站 Twitter. com 上面，很多用户习惯随时发布他们的位置和动态信息，结果有几家网站，如 PleaseRobMe. com（请来抢劫我）、WeKnowYourHouse. com（我知道你的家），能够根据用户所发的信息，推测出用户不在家的时间，找到用户的准确家庭住址，甚至把房子的照片都能找出来。

他们的做法旨在提醒大家我们随时暴露在公众视线下，如果不培养安全和隐私意识，将会给自身带来灾难。目前世界的很多国家，包括中国，都在完善与数据使用及隐私相关的法律，来保护隐私信息不被滥用。

7. 互联互通与数据共享问题

在我国的企业信息化建设过程中，普遍存在条块分割和信息孤岛的现象。不同行业之间的系统与数据几乎没有交集，同一行业，比如交通、社保系统内部等，也是按行政领域进行划分建设，跨区域的信息交互和协同非常困难。严重的甚至在同一单位内，比如一些医院的信息系统建设，病历管理、病床信息、药品管理等子系统都是分立建设的，没有实现信息共享和互通。

“智慧城市”是我国十二五信息化建设的重点，而智慧城市的根本，是要实现信息的互联互通和数据共享，基于数据融合实现智能化的电子政务、社会化管理和民生改善。因此在城市数字化的基础上，还需实现互联化，打通各行各业的数据接口，实现互联互通，在此之上才能实现智慧化。比如在城市应急管理方面，就需要交通、人口、公安、消防、医疗卫生等各个方面的数据和协助。当前美国联邦政府建立的数据共享平台 www. data. gov，我国北京市政府数据资源网（www. bjdata. gov. cn）等都是朝着数据开放、数据共享的有力的尝试。

为实现跨行业的数据整合，需要制定统一的数据标准、交换接口以及共享协议，这样不同行业、不同部门、不同格式的数据才能基于一个统一的基础进行访问、交换和共享。

对于数据访问，还需制定细致的访问权限，规定什么样的用户在什么样的场景下，可以访问什么类型的数据。在大数据及云计算时代，不同行业、企业的数据可能存放在统一的平台和数据中心之上，需要对一些敏感信息进行保护，比如涉及企业商业机密及交易信息方面的数据，虽然是依托平台来进行处理，但是除了企业自身的授权人员之外，要保证平台管理员以及其他企业都不能访问此类数据。

8.1.2　使用大数据分析安全与隐私的问题

曾经有程序员使用 WiFi 登录脚本扫描 WiFi 密码数据，然后对扫描数据做了简单的分析，以便侦测中国家庭 WiFi 通用密码 TOP10。在整个扫描过程中，所有常见的密码和排名的比例保持稳定。因此，它可以是一个基本的判断，表 8.2 是更准确的统计概率。

表 8.2　中国家用 WiFi 常见密码 TOP10

排名	密码	数量	占比	累计占比
1	12345678	3048	3.256%	3.256%
2	123456789	2460	2.628%	5.885%
3	88888888	1453	1.552%	7.437%
4	1234567890	711	0.760%	8.197%
5	00000000	406	0.434%	8.631%
6	87654321	351	0.375%	9.006%
7	66668888	335	0.358%	9.363%
8	11223344	316	0.338%	9.701%
9	147258369	313	0.334%	10.035%
10	11111111	299	0.319%	10.355%

该清单如下：

从列表中，前三名的密码是 12345678、123456789 和 88888888。这三个密码的总数是 7.437%。不要低估了这 7.437%，实际上，这已经被认为非常可怕的比例。因为在九个 WiFi 信号的情况下，WiFi 能够突破这三个密码的概率为 50.1%！请记下这三个密码，以后永远不要使用它们。

我们再往下看，TOP10 的密码列表涵盖所有 WiFi 密码样本的 10.355%。所以中国家庭 WiFi 的安全形势依然十分严峻。

想知道为什么你的账号密码总是被盗吗？研究解密千万密码的背后的密码心理学：设定什么样的密码最安全？数据基因公司的研究结果显示，1234 为最常用密码。

8.2 大数据安全与隐私保护关键技术

8.2.1 基于大数据的威胁发现技术

由于大数据分析技术的出现，企业可以超越以往的“保护－检测－响应－恢复”(PDRR)模式，更主动地发现潜在的安全威胁。例如，IBM 推出了名为 IBM 大数据安全智能的新型安全工具，可以利用大数据来侦测来自企业内外部的安全威胁，包括扫描电子邮件和社交网络，标示出明显心存不满的员工，提醒企业注意，预防其泄露企业机密。

“棱镜”计划也可以被理解为应用大数据方法进行安全分析的成功故事。通过收集各个国家各种类型的数据，利用安全威胁数据和安全分析形成系统方法发现潜在危险局势，在攻击发生之前识别威胁。

相比于传统技术方案，基于大数据的威胁发现技术具有以下优点。

1. 分析内容的范围更大

传统的威胁分析主要针对的内容为各类安全事件。而一个企业的信息资产则包括数据资产、软件资产、实物资产、人员资产、服务资产和其他为业务提供支持的无形资产。由于传统威胁检测技术的局限性，其并不能覆盖这六类信息资产，因此所能发现的威胁也是有限的。

而通过在威胁检测方面引入大数据分析技术，可以更全面地发现针对这些信息资产的攻击。例如通过分析企业员工的即时通信数据、E-mail 数据等可以及时发现人员资产是否面临其他企业“挖墙脚”的攻击威胁。再比如通过对企业的客户部订单数据的分析，也能够发现一些异常的操作行为，进而判断是否危害公司利益。可以看出，分析内容范围的扩大使得基于大数据的威胁检测更加全面。

2. 分析内容的时间跨度更长

现有的许多威胁分析技术都是内存关联性的，也就是说，实时收集数据，采用分析技术发现攻击。分析窗口通常受限于内存大小，无法应对持续性和潜伏性攻击。而引入大数据分析技术后，威胁分析窗口可以横跨若干年的数据，因此威胁发现能力更强，可以有效应对 APT 类攻击。

3. 攻击威胁的预测性

传统的安全防护技术或工具大多是在攻击发生后对攻击行为进行分析和归类，并做出响应。而基于大数据的威胁分析，可进行超前的预判。它能够寻找潜在的安全威胁，对未发生的攻击行为进行预防。

4. 对未知威胁的检测

传统的威胁分析通常是由经验丰富的专业人员根据企业需求和实际情况展开，然而这种威胁分析的结果在很大程度上依赖于个人经验。同时，分析所发现的威胁也是已知的。而大数据分析的特点是侧重于普通的关联分析，而不侧重因果分析，因此通过采用恰当的分析模型，可发现未知威胁。

虽然基于大数据的威胁发现技术具有上述的优点，但是该技术目前也存在一些问题和挑战，主要集中在分析结果的准确程度上。一方面，大数据的收集很难做到全面，而数据又是分析的基础，它的片面性往往会导致分析出的结果的偏差。为了分析企业信息资产面临的威胁，不但要全面收集企业内部的数据，还要对一些企业外的数据进行收集，这些在某种程度上是一个大问题。另一方面，大数据分析能力的不足影响威胁分析的准确性。例如，纽约投资银行每秒会有 5000 次网络事件，每天会从中捕捉 25TB 数据。如果没有足够的分析能力，要从如此庞大的数据中准确地发现极少数预示潜在攻击的事件，进而分析出威胁是几乎不可能完成的任务。

8.2.2 基于大数据的认证技术

身份认证是信息系统或网络中确认操作者身份的过程。传统的认证技术主要通过用户所知的秘密（例如口令），或者持有的凭证（例如数字证书），来鉴别用户。这些技术面临着如下两个问题：

首先，攻击者总是能够找到方法来骗取用户所知的秘密，或窃取用户持有的凭证，从而通过认证机制的认证。例如攻击者利用钓鱼网站窃取用户口令，或者通过社会工程学方式接近用户，直接骗取用户所知秘密或持有的凭证。

其次，传统认证技术中认证方式越安全往往意味着用户负担越重。例如，为了加强认证安全，而采用的多因素认证.用户往往需要同时记忆复杂的口令，还要随身携带硬件 USBKey。一旦忘记口令或者忘记携带 USBKey，就无法完成身份认证。为了减轻用户负担，一些生物认证方式出现，利用用户具有的生物特征，例如指纹等，来确认其身份。然而，这些认证技术要求设备必须具有生物特征识别功能，例如指纹识别。因此在很大程度上限制了这些认证技术的广泛应用。

而在认证技术中引入大数据分析则能够有效地解决这两个问题。基于大数据的认证技术指的是收集用户行为和设备行为数据，并对这些数据进行分析，获得用户行为和设备行为的特征，进而通过鉴别操作者行为及其设备行为来确定其身份。这与传统认证技术利用用户所知秘密、所持有凭证或具有的生物特征来确认其身份有很大不同。具体来说，这种新的认证技术具有如下优点：

（1）攻击者很难模拟用户行为特征来通过认证，因此更加安全。利用大数据技术所能收集的用户行为和设备行为数据是多样的，可以包括用户使用系统的时间、经常采用的设备、设备所处物理位置，甚至是用户的操作习惯数据。通过这些数据的分析能够为用户勾画一个行为特征的轮廓。而攻击者很难在方方面面都模仿出用户的行为，因此其与真正用户的行为特征轮廓必然存在一个较大偏差，无法通过认证。

（2）减小了用户负担.用户行为和设备行为特征数据的采集、存储和分析都由认证系统完成。相比于传统认证技术，极大地减轻了用户负担。

（3）可以更好地支持各系统认证机制的统一，基于大数据的认证技术可以让用户在整个网络空间采用相同的行为特征进行身份认证，而避免不同系统采用不同认证方式，且用户所知秘密或所持有凭证也各不相同而带来了种种不便。

虽然基于大数据的认证技术具有上述优点，但同时也存在一些问题和挑战亟待解决：

(1) 初始阶段的认证问题。

基于大数据的认证技术是建立在大量用户行为和设备行为数据分析的基础上,而初始阶段不具备大量数据。因此,无法分析出用户行为特征,或者分析的结果不够准确。

(2) 用户隐私问题。

基于大数据的认证技术为了能够获得用户的行为习惯,必然要长期持续地收集大量的用户数据。那么如何在收集和分析这些数据的同时,确保用户隐私也是亟待解决的问题。它是影响这种新的认证技术是否能够推广的主要因素。

8.2.3 基于大数据的数据真实性分析

目前,基于大数据的数据真实性分析被广泛认为是最为有效的方法。许多企业已经开始了这方面的研究工作,例如 Yahoo 和 Thinkmail 等利用大数据分析技术来过滤垃圾邮件;Yelp 等社交点评网络用大数据分析来识别虚假评论;新浪微博等社交媒体利用大数据分析来鉴别各类垃圾信息等。

基于大数据的数据真实性分析技术能够提高垃圾信息的鉴别能力。一方面,引入大数据分析可以获得更高的识别准确率。例如,对于点评网站的虚假评论,可以通过收集评论者的大量位置信息、评论内容、评论时间等进行分析,鉴别其评论的可靠性。如果某评论者为某品牌多个同类产品都发表了恶意评论,其评论的真实性就值得怀疑。另一方面,在进行大数据分析时,通过机器学习技术,可以发现更多具有新特征的垃圾信息。然而该技术仍然面临一些困难,主要是虚假信息的定义、分析模型的构建等。

8.2.4 大数据与“安全即服务”

前面列举了部分当前基于大数据的信息安全技术,未来必将涌现出更多、更丰富的安全应用和安全服务。由于此类技术以大数据分析为基础,因此如何收集、存储和管理大数据就是相关企业或组织所面临的核心问题。除了极少数企业有能力做到之外,对于绝大多数信息安全企业来说,更为现实的方式是通过某种方式获得大数据服务,结合自己的技术特色领域,对外提供安全服务。一种未来的发展前景是:以底层大数据服务为基础,各个企业之间组成相互依赖、相互支撑的信息安全服务体系,从总体上形成信息安全产业界的良好生态环境。

8.3 大数据安全的防护策略

1. 确保身份安全

要进行大数据分析,需要把大型数据集划分成更易于管理的单个部分,然后分别通过 Hadoop 集群处理,最后将它们重新组合以产生所需分析。该过程高度自动化,涉及大量跨集群的机器对机器(M2M)交互。

在 Hadoop 的基础设施会发生几个层次的授权,具体包括:

(1) 访问 Hadoop 集群。

(2) 簇间通信。

(3) 集群访问数据源。

这些授权往往是基于 SSH(Secure Shell)密钥的,其对于使用 Hadoop 是理想的,因其安全级别支持自动化的 M2M 通信。

许多基于流行的基于云计算的 Hadoop 服务也使用 SSH 作为访问 Hadoop 集群的认证方法。确保了授予访问大数据环境中的身份应该是一个高优先级的,但其也具有挑战性。这对于那些想要像使用 Hadoop 一样使用大数据分析的公司来说是一个很大的挑战。有些问题直截了当:

- 谁来建立运行大数据分析的授权?
- 一旦建立授权的人离职,会出现什么问题?
- 授权提供的访问级别是否基于"须知"安全准则?
- 谁可以访问授权?
- 如何管理这些授权?

大数据并不是需要考虑这些问题的唯一技术。当越来越多的业务流程自动化,这些问题将遍布数据中心。自动化的 M2M 交易占到了数据中心所有通信的 80%,然而大部分管理员则把焦点集中在与员工账户相关联的 20%的通信流量。

2. 风险

众所周知的数据泄露包括滥用以机器为主的证书,这体现了忽视 M2M 身份验证的现实风险。当企业在管理终端用户身份上取得很大进步时,却忽视了应以同样标准处理机器为主的身份验证的需求。其结果就是使整个 IT 环境遍布风险。

然而,对于想要将集中的身份和存取管理(尽可能的)应用到数百万基于机器的身份来说,改变运行中的系统是一个很大的挑战。不中断系统迁移环境是一项复杂的工作,所以企业一直在犹豫也不足为奇。

3. 密钥管理的不良状况

密钥管理的现状一直很糟糕。为了管理用于保护 M2M 通信的认证密钥,许多系统管理员使用电子表格或自编脚本来控制分配、监控和清点密钥。这种做法漏掉了许多密钥。估计他们也没有设置常规扫描,于是未被授权的非法途径便在不知不觉中添加进来。

缺少对密钥的集中控制严重影响法规遵从。以金融行业为例,规定要求必须严格控制谁可以访问敏感数据,比如最近强化了的 PCI 标准要求任何接受支付卡的地方——银行、零售商、餐馆和医院等——均需依照同样标准执行,无一例外。由于这些行业目前正在迅速果断地执行大数据战略,来分得用户驱动数据大潮的一杯羹,他们越来越容易违背法规并面临监管制裁。

4. 安全步骤

组织机构必须承认并应对这些风险。这些步骤是行动开始的最佳做法:

- 很少有 IT 人员知道身份的存储位置、访问权限以及其支持的业务流程。因此,第一步是被动非侵入的发现。
- 环境监测是必需的,这样才能确定哪些身份是活跃的,哪些不是。幸运的是,在许多企业中,未使用的——因此也是不需要的——身份往往占绝大多数。一旦这些

未使用的身份被定位并移除，整体工作量便会大大降低。

- 下一步是集中控制添加、更改和删除机器身份。这样一来，政策便可以控制身份如何使用，确保没有非托管的身份添加，并提供法规遵从的有效证明。
- 随着可见性和管理控制的确定，必要但在违反政策的身份可以在不中断业务流程的情况下进行校正。集中管理可对该身份的权限级别进行修正。

5. 安全策略

大数据的兴起伴随着数据存取控制的新型风险。M2M 身份管理必不可少，但是传统的人工 IAM 做法效率低且风险高。盘点所有密钥，使用最优方法可以节省时间和金钱，同时提高安全性和法规遵从。由于大数据增加了访问敏感信息的认证门槛，组织机构必须采取积极措施，推出全面一致的身份和存取管理策略。

8.4 大数据应用案例之：电影《爸爸去哪儿》大卖有前兆么

1 月 25 日《爸爸去哪儿》北京首映发布会，首映广告如图 8.1 所示。有媒体问郭涛：电影拍摄期只有 5 天，你怎么让观众相信，这是一部有品质的电影？



图 8.1 《爸爸去哪儿》

1 月 27 日，某娱乐频道挂出头条策划——只拍了 5 天的《爸爸去哪儿》，值得走进电影院么？

光线传媒总裁王长田在 1 月 7 日的发布会上解释：一般的电影只有 2 或 3 台摄影机，而《爸爸去哪儿》用了 30 多台摄影机，所以这 5 天的拍摄时间，却有 10 倍的素材量，在剪辑量上，甚至比一般电影还大。

关于这个话题的讨论，看上去好像很多很多，但，观众真的很在乎这件事情么？《爸爸去哪儿》大卖，到底是让大家大跌眼镜的偶然事件，还是早有前兆？

1. 5 天拍完，观众真的在乎么

《爸爸去哪儿》是一部真人秀电影，制作流程在中国电影史上也没有可参照对象，但我们依然担心会有很多人贴标签——5 天拍完，粗制滥造！

本着危机公关心态出发，伯乐营销委托微瑞科技做了一次大数据挖掘，想看看到底有多少人在质疑《爸爸去哪儿》圈钱的事情，以及在提到这件事情时，截止到 1 月 11 日，在新

浪微博仅有536人在讨论此话题(含转发人数),而其中还有近一半的人数表示出圈钱也要看、圈钱也无所谓的态度。比起动辄对影片内容数十万的讨论和追捧,这个话题讨论量简直是沧海一粟。

反过来想一想,一部电影拍5年,就值得进电影院了么?在商业电影环境里,让观众买单的是结果,而不是努力的过程。

网友对“圈钱”的看法如图8.2所示。



图8.2　网友对“圈钱”的看法

2. 原班人马出演,很重要么

由热门电视剧改编成的电影的项目有很多,这里面有票房大卖的,比如《武林外传》《将爱》,也有票房惨淡的,比如《奋斗》《宫》《金太狼的幸福生活》,用比较粗暴的方式分类,前者基本是原班人马出演,而后者的主演都变了,虽然单单看这几个项目,就推出“热门电视剧改编电影+原班人马=票房大卖”未免太粗暴,但不得不说,原班人马对于项目成功一定是加分因素。而通过大数据调查发现,对于《爸爸去哪儿》大电影这个项目来说,更是至关重要。

当《爸爸去哪儿》大电影项目刚刚曝光,还没有公布主演的短短两三天时间,在新浪微博上参与“原班人马”的讨论量便已经过超过2万,43.38%的网友表示,如果是原班人马就会看;47.06%的人表示,希望是原班人马出演;9.56%的网友表示,不是原班人马不看。无论是从声量的绝对值上,还是从期待人数的比例上,都能够证明五对星爸萌娃合体的价值。

原班人马出演很重要的分析统计如图8.3所示。

与此同时,我们也调研了《武林外传》《将爱》《宫》《奋斗》四部电影在新浪微博上对于“原班人马”的讨论,虽然样本比《爸爸去哪儿》大电影小得多,但从分布比例上,能明显看出,《奋斗》和《宫锁沉香》没有使用原班人马,有失人心。

失败案例如图8.4所示。

3. 谁才是真正的“合家欢”

今年春节档公映的电影,几乎每一部电影在宣传时,都要给自己贴上“合家欢”电影的标签,“合家欢”真的那么重要么?

每年的春节档都是电影院的业务爆发期,而在这其中最受电影院欢迎的电影就是合家欢类型的电影,不难理解,适合全家一起看的,能够带动更多消费,这样类型的电影必然受电影院的欢迎。以春节档的《大闹天宫》《爸爸去哪儿》《澳门风云》作为调研对象,在新

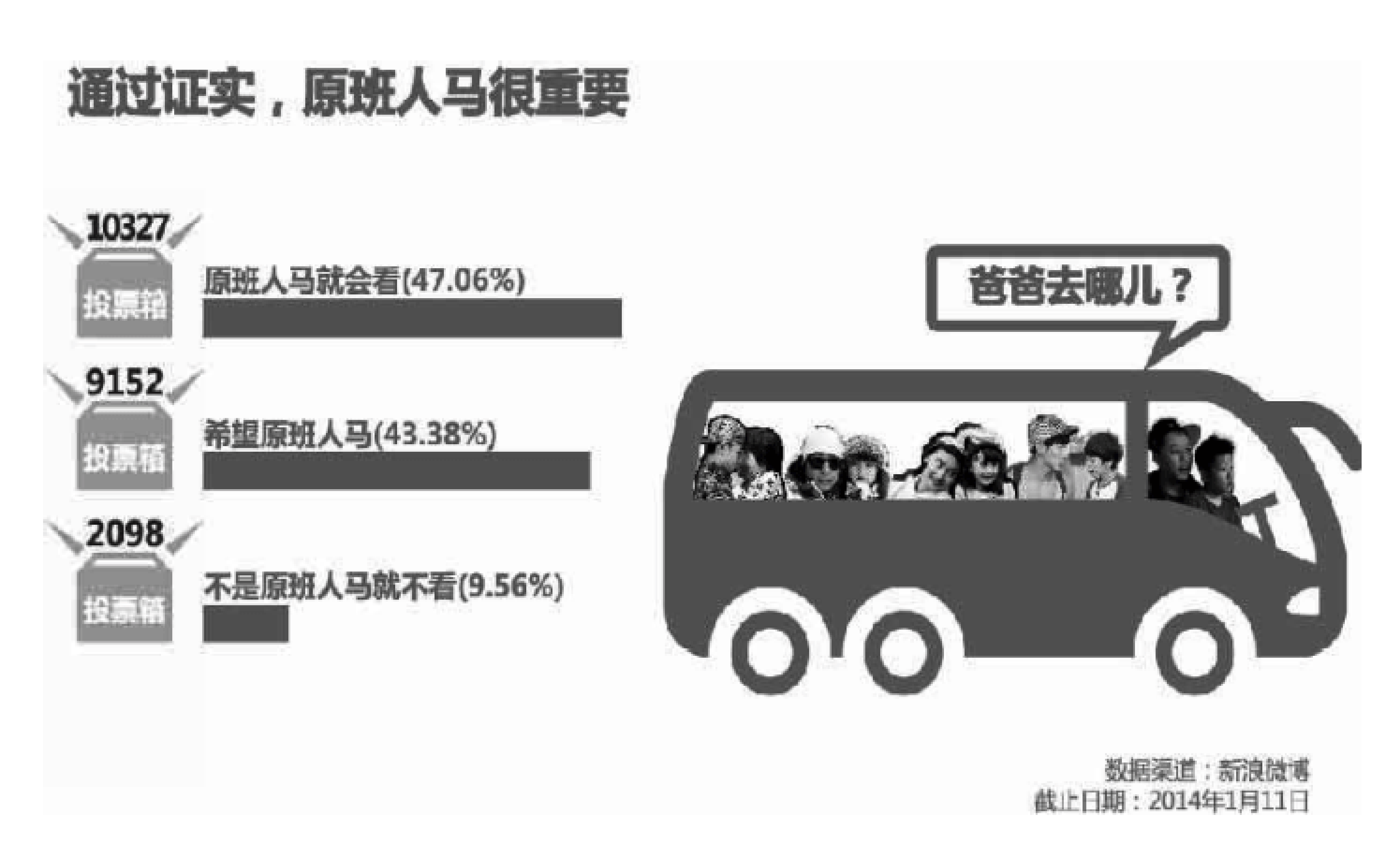


图 8.3　原班人马出演，很重要吗？

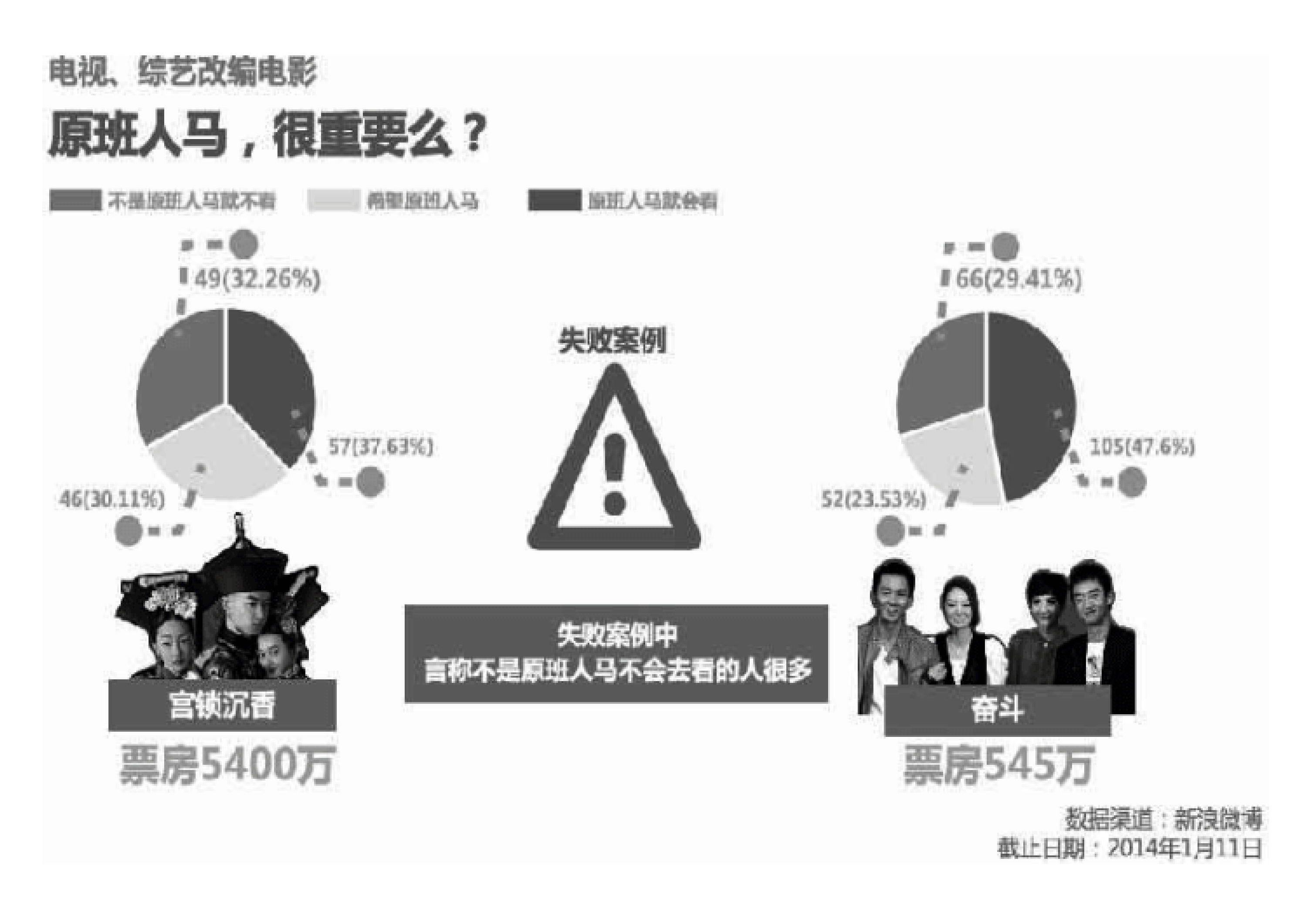


图 8.4　失败案例

浪微博上抽取以电影名和全家为关键词为讨论的条目，相较于《澳门风云》《爸爸去哪儿》和《大闹天宫》有着绝对的优势。

4. 预告片播放量，你有注意么

数据公司随机抽取了四部在大年初一公映的四部电影在 12 月以来发布的一款预告片和一款制作特辑，以腾讯、搜狐、新浪、优酷、土豆五家主流视频网站的播放量为调研对象，可以明显看出《爸爸去哪儿》和《大闹天宫》都是百万量级播放量，在四部电影里占尽优势。（因为《大闹天宫》项目启动较早，考虑到预告片在各个平台上的长尾效应，并没有选择《大闹天宫》一首款预告，而是选择了与其他电影在密集宣传期时相近时间主推的预告片。）

预告片播放量的统计分析如图 8.5 所示。

5. 春节档，到底最想看什么

其实论来论去，片方最紧张的还是“想看”一部电影的人数，毕竟“想看”这个词，直接和票房挂钩，但这个问题最复杂，因为提供这个数据的维度非常多，有新浪微博上网友直

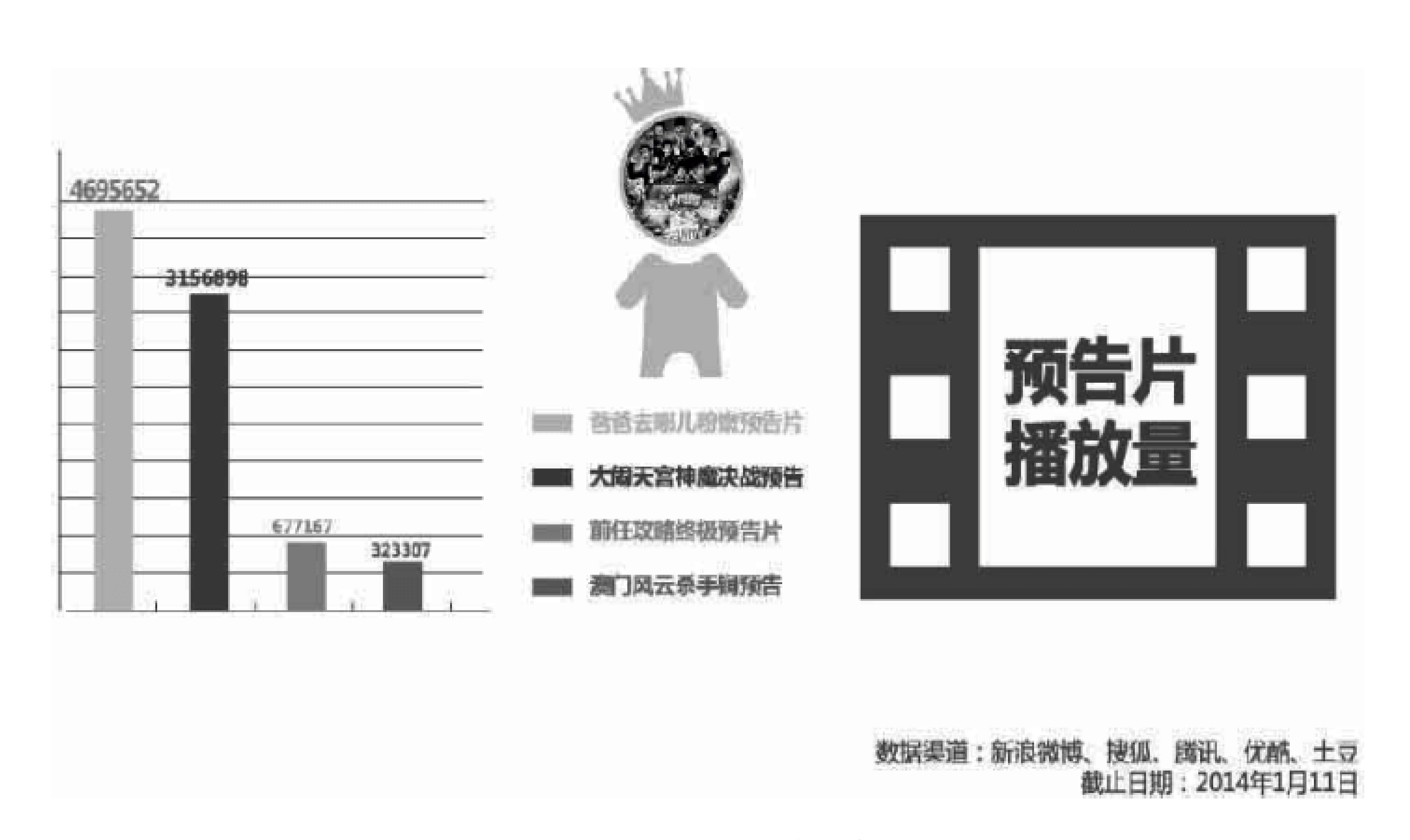


图 8.5　预告片播放量

接发出的声音,有业内营销人士非常关心的百度指数,也有像 QQ 电影票这样和用户购买行为直接相关的 APP,所以在这个问题上,我们特地选择了几个不同的平台取样。

1）新浪微博

从图 8.6 来看,在大年初一即将上映的四部电影中,提及“想看”和“期待”《前任攻略》的频率最高,《爸爸去哪儿》电影次之,这个结果可能和大家对未来各个电影在票房上的期待不符,但它的确反映了在微博这个阵地上各个电影因为粉丝而带来的话题讨论量。

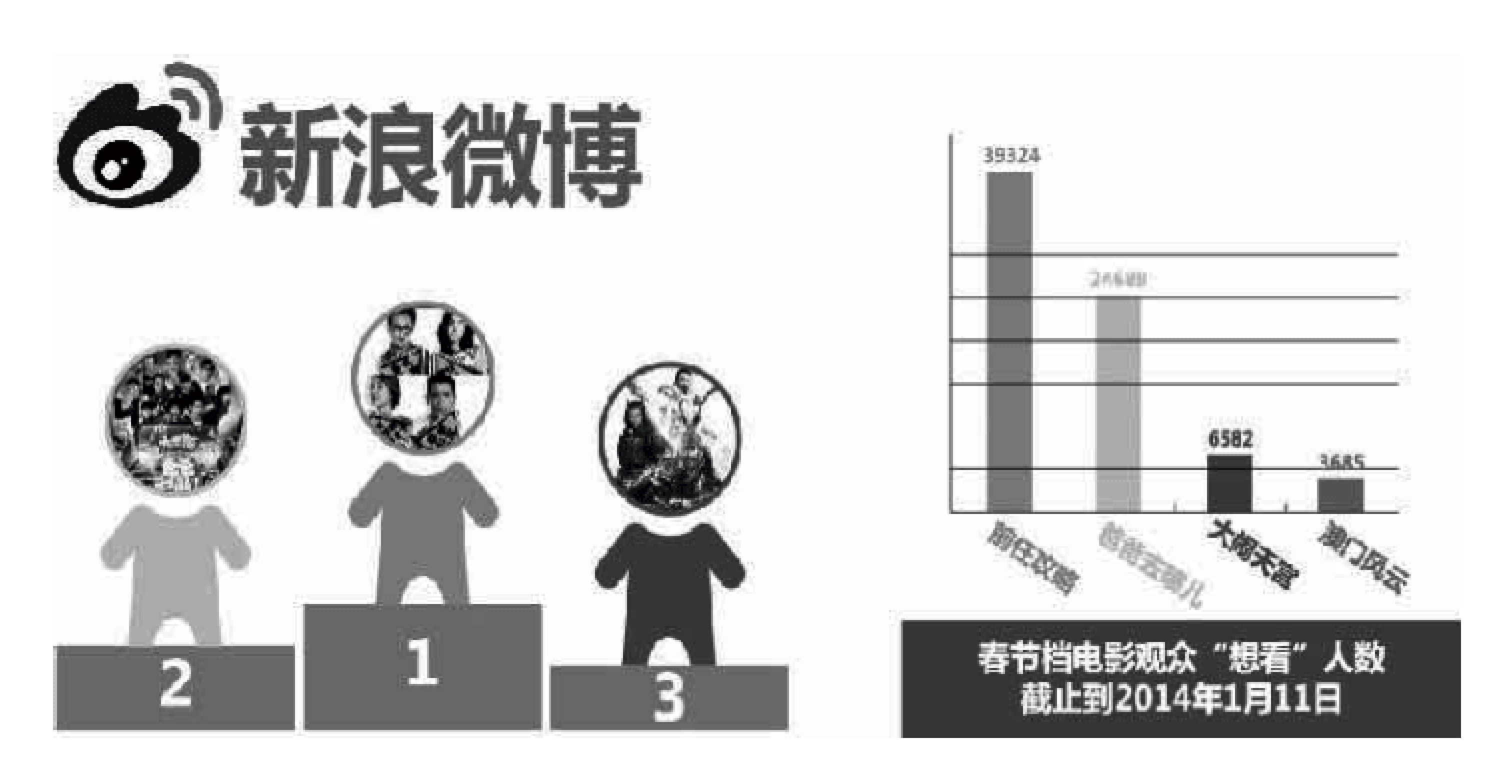


图 8.6　新浪微博数据图

2）百度指数

百度指数是电影从业人员比较看中的一个数据,之前还有人以此建模,预测电影最终票房,这个数据代表每天有多少人以片名为关键词进行搜索,在某种程度上,它的确可以反映出一部电影在网民心中的热度。因此,数据公司选取了四部在大年初一上映的电影的百度指数在 1 月 26 日之前映前 30 天的平均值作为参考。

需要说明的是——由于《爸爸去哪儿》这部电影有综艺节目的干扰,关键词选择了“爸爸去哪儿电影”,虽然这样会漏掉大量搜索数据(如此期间搜索“爸爸去哪儿大电影”的平均指数,也达到了 4200,但都没有统计在内),但即便如此,《爸爸去哪儿》的平均指数仍在 4 部电影里排名第 2。

百度指数数据分析如图 8.7 所示。

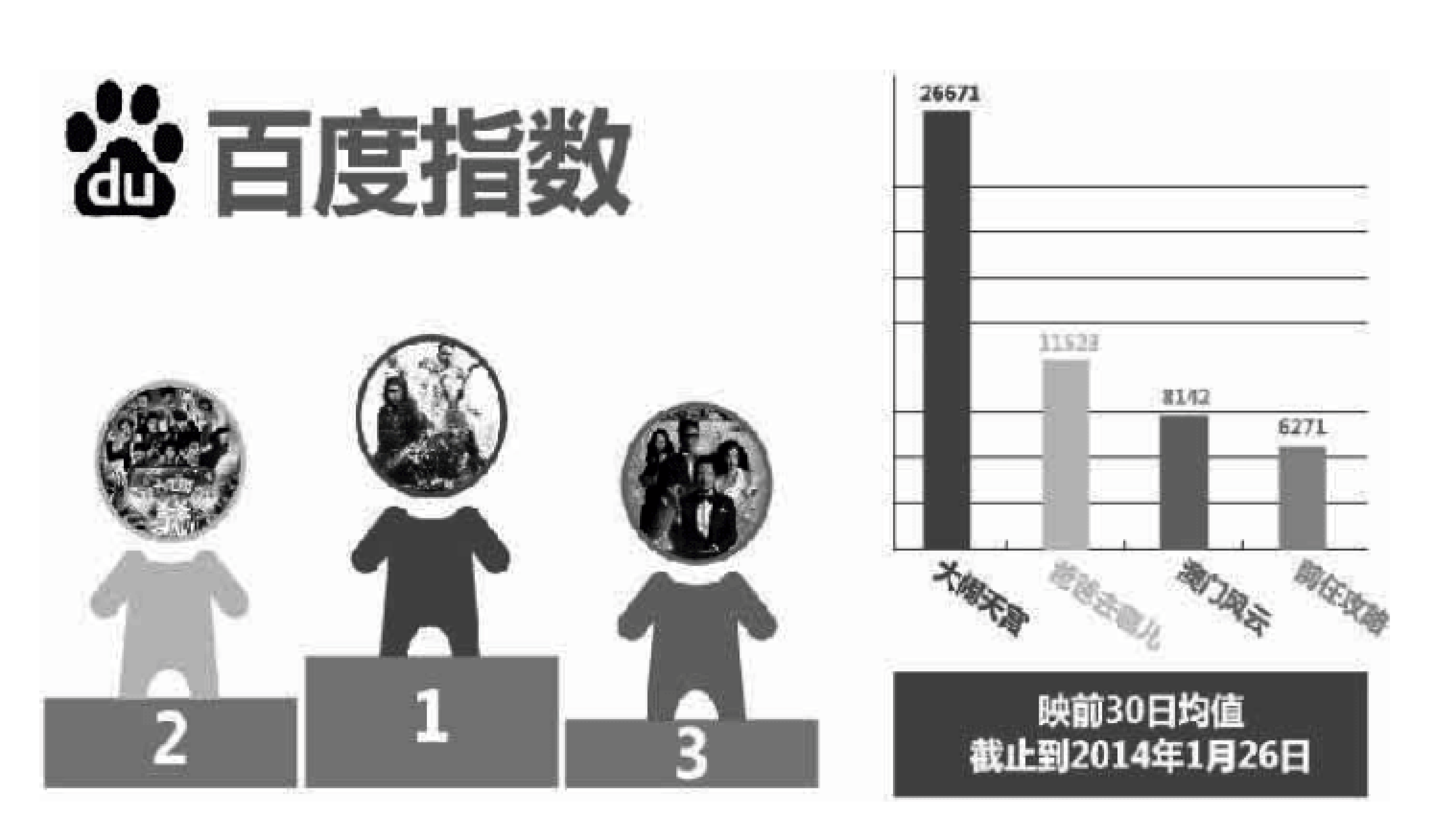


图 8.7 百度指数数据图

3）猫眼电影、QQ 电影票

美团猫眼电影和 QQ 电影票是当下两款非常流行的购票服务软件，提供影票查询信息、团购影票，以及在线座位提早预订等功能，其 APP 软件已经成为学生和白领一族购买电影票的重要渠道，而在这个渠道上所体现出的“想看”和最后的消费购买距离最近，从这两个数据上来看，《大闹天宫》和《爸爸去哪儿》电影的优势最为明显。

值得注意的是，《爸爸去哪儿》电影项目公布的时间是这 4 个项目里最晚的一部，在 12 月初时，当其他几部电影已经有了一个不错的基数时，《爸爸去哪儿》还是 0，但就在这 1 个多月里，“想看”的数字形成了一个爆发式的增长。不过这两个平台都不提供体现变化的数据，所以大家看不到这部电影在“想看”这个数据上突飞猛进的变化。

使用购票服务软件分析结果如图 8.8 所示。

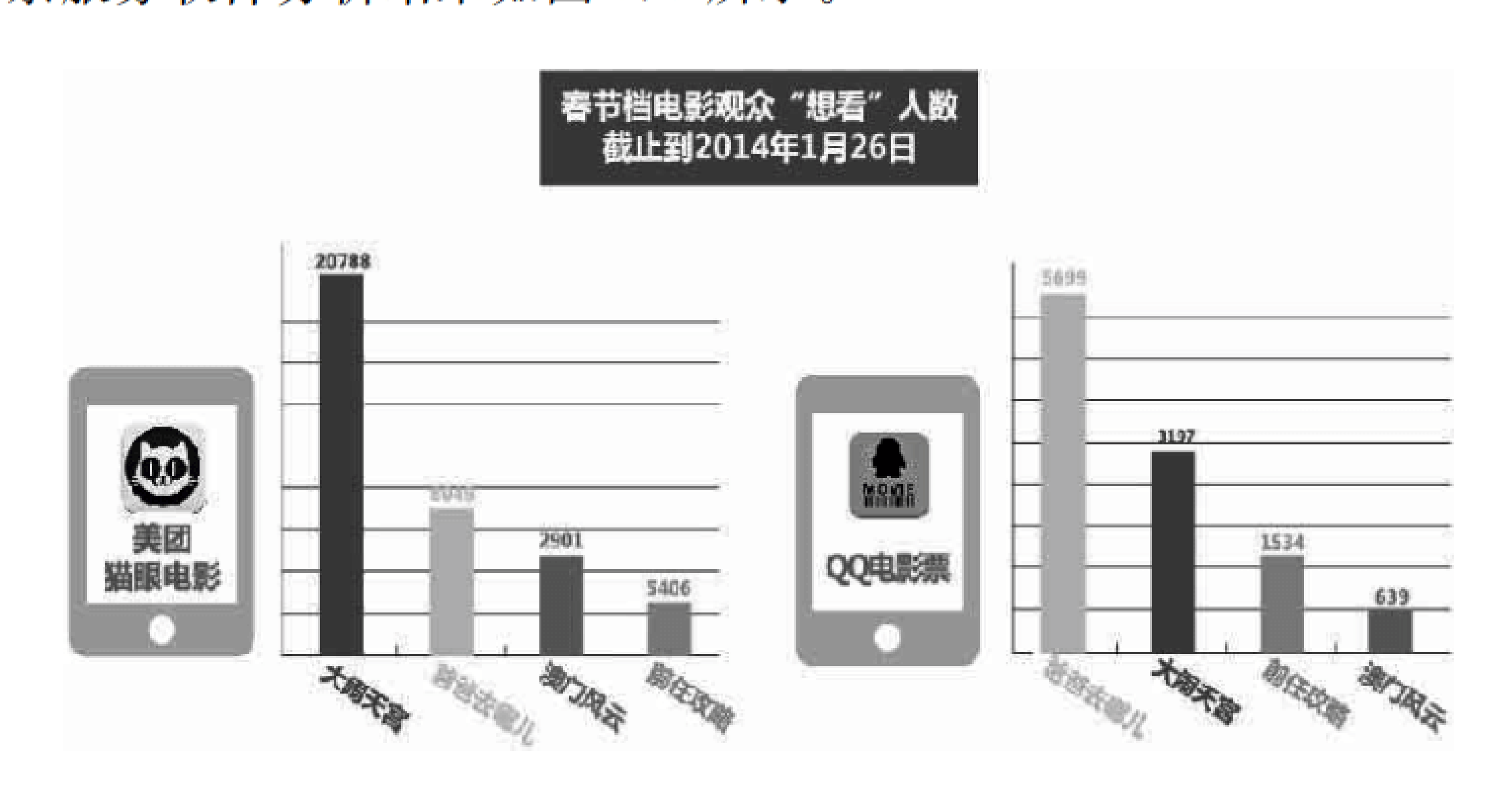


图 8.8 猫眼电影、QQ 电影票数据图

6. 微博营销，谁的影响力最大

每逢片方发布新的预告片、海报、特辑等重要物料，都会与演员沟通能在其微博上配合发布，作为一个动辄上百万，甚至上千万粉丝的明星，微博的确是一个非常有效的话题和内容的输出渠道，其效果在《小时代》和《致青春》这两部电影上更是有着现象级的释放。针对微博营销，数据公司选取了 12 月 26 日～1 月 26 日这个时间段里，四部电影的主演在微博上配合发布电影物料的总条数，以及由此带来的转发量、评论量。

在转发量、评论量上，《爸爸去哪儿》代表队都以绝对的优势成为最大赢家，其中林志颖共计发布 12 条微博，转发量 20.2 万，评论量 17.8 万，毫无悬念地成为称霸微博人气王，而田亮发布的 13 条微博，带来了 10.1 万转发、6.8 万的评论量，影响力也不容小觑，《大闹天宫》代表队的主力选手是主演甄子丹，期间共发布 58 条微博，可以说是名副其实的互动王，《前任攻略》代表队中影响力最大的是韩庚，《澳门风云》在微博平台上相对比较吃亏，因为电影里的两位重要级演员周润发、谢霆锋都没有开通新浪微博，期间大部分与网友的互动，都是由导演王晶完成。

除了演员带来的巨大影响力之外，《爸爸去哪儿》综艺节目而建立起来的官方微博，也成为《爸爸去哪儿》电影版后期的宣传平台，400 余万的粉丝数量，也是其他两三万粉丝数的电影官微所不能比的，所以在物料和新媒体话题的传播上，《爸爸去哪儿》占尽了优势。

微博营销影响力数据分析如图 8.9 所示。

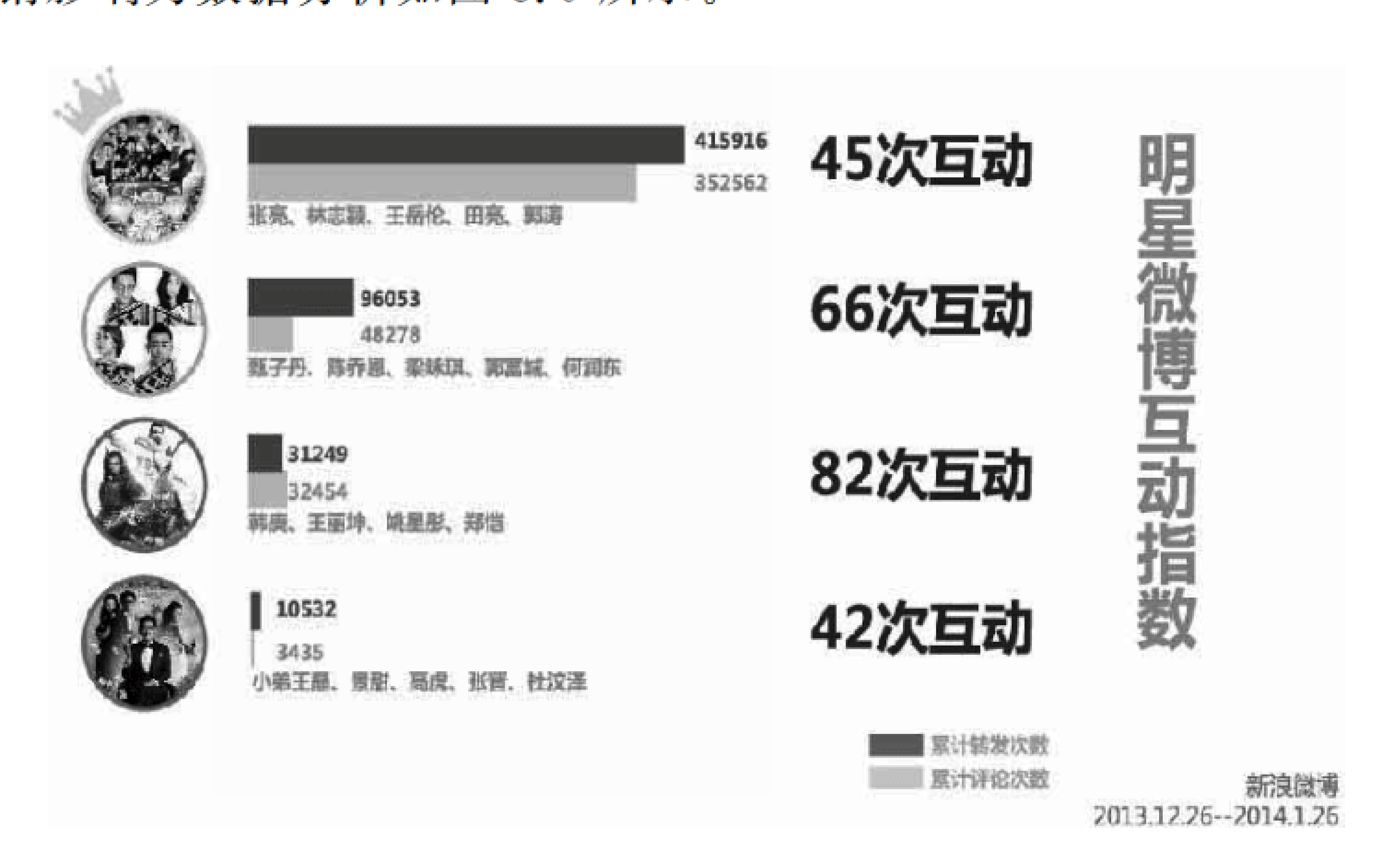


图 8.9　微博营销影响力数据图

作为中国首部真人秀电影，《爸爸去哪儿》有它的独创性，也有它的不可复制性，也许它不是好的艺术范本，但它一定是一个好的商业范本，如果你还在拿它与《中国好声音之为你转身》相比，那真是把这个项目想得太简单了。在时间点的衔接、电影的内容、档期的安排、台网宣传互动、联合营销推广方面，虽然项目启动最晚，但《爸爸去哪儿》在上述各个方面，都有着精心的安排和规划，而其之前在社交媒体上强劲的数据表现也证明，《爸爸去哪儿》能大卖，真的没什么可意外的。

习题与思考题

一、选择题

1. 以下哪些管理规定对信息安全及个人隐私进行了保护？(　　)

A.《互联网行业的自律公约》　　B.《治安管理处罚条例》

C.《关于加强网络信息保护的决定》　　D.《信息安全保护条例》

2. 在大数据时代，我们需要设立一个不一样的隐私保护模式，这个模式应该更着重于(　　)为其行为承担责任。

A. 数据使用者　　B. 数据提供者
C. 个人许可　　D. 数据分析者

3. 云安全主要的考虑的关键技术有哪些？（　　）
A. 数据安全　　B. 应用安全
C. 虚拟化安全　　D. 服务器安全

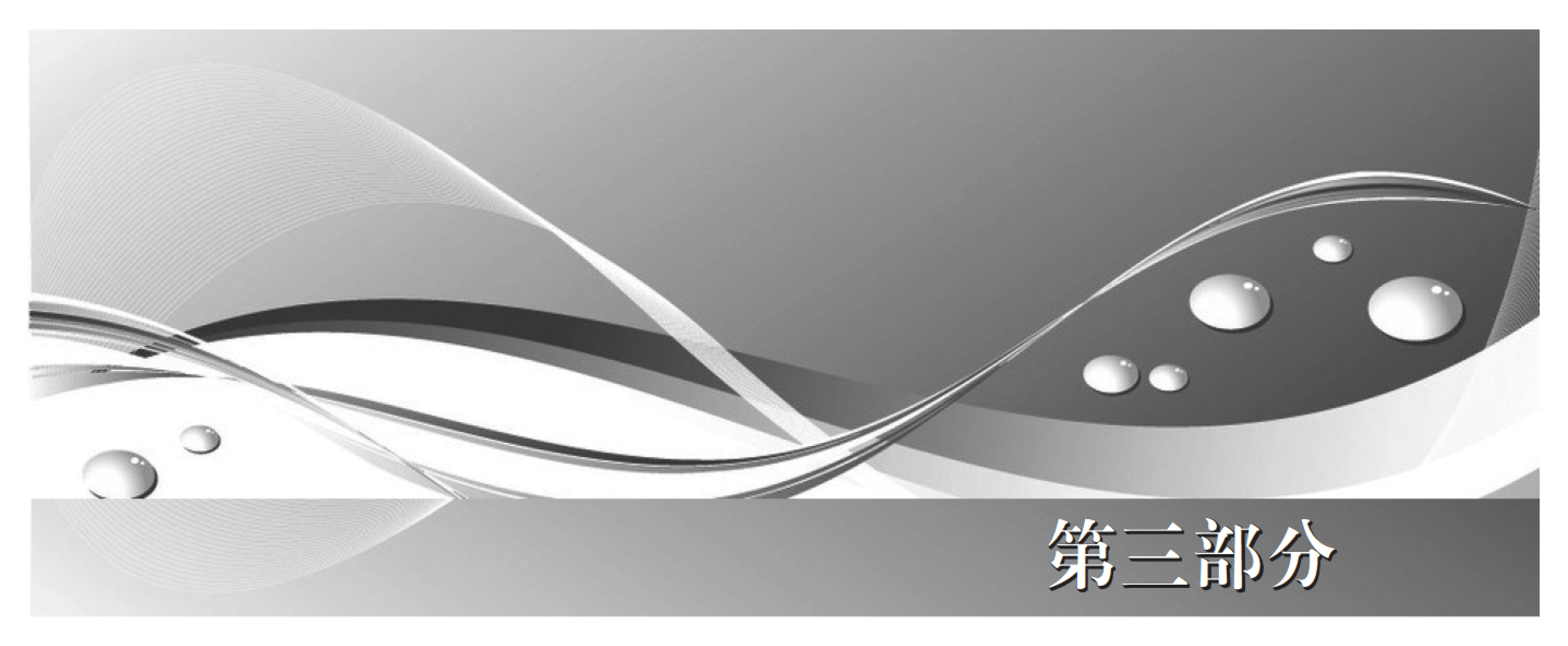
4. 下列关于网络用户行为的说法中，错误的是（　　）。
A. 网络公司能够捕捉到用户在其网站上的所有行为
B. 用户离散的交互痕迹能够为企业提升服务质量提供参考
C. 数字轨迹用完即自动删除
D. 用户的隐私安全很难得以规范保护

5. 下列论据中，能够支撑“大数据无所不能”的观点的是（　　）。
A. 互联网金融打破了传统的观念和行为
B. 大数据存在泡沫
C. 大数据具有非常高的成本
D. 个人隐私泄露与信息安全担忧

6. 促进隐私保护的一种创新途径是（　　）：故意将数据模糊处理，促使对大数据库的查询不能显示精确的结果。
A. 匿名化　　B. 信息模糊化
C. 个人隐私保护　　D. 差别隐私

二、问答题

1. 大数据面临哪些方面的安全问题？
2. 简述基于大数据的威胁发现技术。
3. 有哪些种基于大数据的认证技术？
4. 简述大数据安全的防护策略。



大数据分析案例

第 9 章　行业案例研究
——银行、保险、证券、金融行业

9.1　银行业应用

9.1.1　大数据时代：银行如何玩转数据挖掘

银行信息化的迅速发展，产生了大量的业务数据。从海量数据中提取出有价值的信息，为银行的商业决策服务，是数据挖掘的重要应用领域。汇丰、花旗和瑞士银行是数据挖掘技术应用的先行者。如今，数据挖掘已在银行业有了广泛深入的应用。

现阶段，数据挖掘在银行业中的应用，主要可分为以下几个方面。

1. 风险控制

数据挖掘在银行业的重要应用之一是风险管理，如信用风险评估。可通过构建信用评级模型，评估贷款人或信用卡申请人的风险。一个进行信用风险评估的解决方案，能对银行数据库中所有的账户指定信用评级标准，用若干数据库查询就可以得出信用风险的列表。这种对于高/低风险的评级或分类，是基于每个客户的账户特征，如尚未偿还的贷款、信用调降报告记录、账户类型、收入水平及其他信息等。

对于银行账户的信用评估，可采用直观量化的评分技术。将顾客的海量信息数据以某种权重加以衡量，针对各种目标给出量化的评分。以信用评分为例，通过由数据挖掘模型确定的权重，来给每项申请的各项指标打分，加总得到该申请人的信用评分情况。银行根据信用评分来决定是否接受申请，确定信用额度。过去，信用评分的工作由银行信贷员完成，只考虑几个经过测试的变量，如就业情况、收入、年龄、资产、负债等。现在应用数据挖掘的方法，可以增加更多的变量，提高模型的精度，满足信用评价的需求。

通过数据挖掘，还可以异常的信用卡使用情况，确定极端客户的消费行为。根据历史数据，评定造成信贷风险客户的特征和背景，可能造成风险损失的客户。在对客户的资信和经营预测的基础上，运用系统的方法对信贷风险的类型和原因进行识别、估测，发现引起贷款风险的诱导因素，有效地控制和降低信贷风险的发生。通过建立信用欺诈模型，帮助银行发现具有潜在欺诈性的事件，开展欺诈侦查分析，预防和控制资金非法流失。

2. 客户管理

在银行客户管理生命周期的各个阶段，都会用到数据挖掘技术。

1）获取客户

发现和开拓新客户对任何一家银行来说都至关重要。通过探索性的数据挖掘方法，

如自动探测聚类和购物篮分析,可以用来找出客户数据库中的特征,预测对于银行活动的响应率。那些被定为有利的特征可以与新的非客户群进行匹配,以增加营销活动的效果。

数据挖掘还可从银行数据库存储的客户信息中,可以根据事先设定的标准找到符合条件的客户群,也可以把客户进行聚类分析让其自然分群,通过对客户的服务收入、风险等相关因素的分析、预测和优化,找到新的可赢利目标客户。

2)保留客户

通过数据挖掘,在发现流失客户的特征后,银行可以在具有相似特征的客户未流失之前,采取额外增值服务、特殊待遇和激励忠诚度等措施保留客户。比如,使用信用卡损耗模型,可以预测哪些客户将停止使用银行的信用卡,而转用竞争对手的卡,根据数据挖掘结果,银行可以采取措施来保持这些客户的信任。当得出可能流失的客户名单后,可对客户进行关怀访问,争取留住客户。

为留住老客户,防止客户流失,就必须了解客户的需求。数据挖掘,可以识别导致客户转移的关联因子,用模式找出当前客户中相似的可能转移者,通过孤立点分析法可以发现客户的异常行为,从而使银行避免不必要的客户流失。数据挖掘工具,还可以对大量的客户资料进行分析,建立数据模型,确定客户的交易习惯、交易额度和交易频率,分析客户对某个产品的忠诚程度、持久性等,从而为他们提供个性化定制服务,以提高客户忠诚度。

3)优化客户服务

银行业竞争日益激烈,客户服务的质量是关系到银行发展的重要因素。客户是一个可能根据年费、服务、优惠条件等因素而不断流动的团体,为客户提供优质和个性化的服务,是取得客户信任的重要手段。根据二八原则,银行业 20%的客户创造了 80%的价值,要对这 20%的客户实施最优质的服务,前提是发现这 20%的重点客户。重点客户的发现通常是由一系列的数据挖掘来实现的。如通过分析客户对产品的应用频率、持续性等指标来判别客户的忠诚度,通过交易数据的详细分析来鉴别哪些是银行希望保留的客户。找到重点客户后,银行就能为客户提供有针对性的服务。

3. 数据挖掘在银行业的具体应用

数据挖掘技术在银行业中的应用,其中一个重要前提条件是,必须建立一个统一的中央客户数据库,以提高客户信息的分析能力。分析开始时,从数据库中收集与客户有关的所有信息、交易记录,进行建模,对数据进行分析,对客户将来的行为进行预测。具体应用分为五个阶段:

(1) 加载客户账号信息。这一阶段,主要是进行数据清理,消除现有业务系统中有关客户账户数据不一致的现象,将其整合到中央客户信息库。银行各业务部门对客户有统一的视图,可以进行相关的客户分析,如客户人数、客户分类、基本需求等。

(2) 加载客户交易信息阶段。这一阶段主要是把客户与银行分销渠道的所有交易数据,包括柜台、ATM、信用卡、汇款、转账等,加载到中央市场客户信息库。这一阶段完成后,银行可以分析客户使用分销渠道的情况和分销渠道的容量,了解客户、渠道,服务三者之间的关系。

(3) 模型评测。这是为客户的每一个账号建立利润评测模型,需要收入和的确定金

额，因此需要加载系统的数据到中央数据库。这一阶段完成后，银行可以从组织、用户和产品三个方面分析利润贡献度。如银行可以依客户的利润贡献度安排合适的分销渠道，模拟和预测新产品对银行的利润贡献度等。

(4) 优化客户关系。银行应该掌握客户在生活、职业等方面的行为变化及外部的变化，抓住推销新产品和服务的时机。这需要将账号每天发生的交易明细数据，定时加载到中央数据仓库，核对客户行为的变化。如有变化，银行则利用客户的购买倾向模型、渠道喜好模型、利润贡献模型、信用和风险评测模型等，主动与客户取得联系。

(5) 风险评估。银行风险管理的对象主要是与资产和负债有关的风险，因此与资产负债有关的业务系统的交易数据要加载到中央数据仓库；然后，银行应按照不同的期间，分析和计算利率敏感性资产和负债之间的缺口，知道银行在不同期间资本比率、资产负债结构、资金情况和净利息收入的变化。

9.1.2　工商银行客户关系管理案例

传统银行的转型实战：看工商银行如何利用大数据洞察客户心声?

1. 工商银行文本挖掘技术应用探索分享

工商银行在大家传统的印象当中是一个体形非常庞大但是稳步前行的形象，但是近些年来在大数据的挑战下工商银行积极应对外界变化，做了一些转型。其中一个举措就是通过数据应用驱动业务变革。

工商银行每天都在面临着来自各方的海量的客户心声，最近 95588 接到这样一个来电，李先生做了一笔跨行汇款操作，对方还没有收到，他来询问什么时候可以到账，这是一个典型的咨询电话。客户王先生是一个贵宾客户，他来电反映说在机场和火车站没有享受到工行提供的贵宾厅，他希望工行在以上场所做明显提示。还有张小姐到一个支行网点做存款业务，发现里面柜员服务态度不耐烦，让她很不满意，她要求把这个情况记录下来做一个反映，这是一个典型的投诉电话。来自各方的海量数据分析如图 9.1 所示。

除了官方服务渠道之外，现在客户越来越希望通过互联网社交网络的方式表达他们的心声，并探讨热点话题。最近我们监测到这样一个热点话题的讨论，有人说：“大家看清楚了，针孔摄像头就是这样装进 ATM 机偷看你的密码的。”这是一个风险事件，工商银行需要做到及时了解和掌握。

同时在互联网的新闻网站上最近也有一些报道，有的市民在便利店蹭 WiFi，上了两个小时网，他的银行卡就被盗刷了，这个又是怎么办到的? 工商银行需要对这些事件做到了解掌控，并且制定对应的措施。以上这些信息都是以文本方式存在的，我们可以通过文本挖掘的方法了解用户在说什么，挖掘出对我们有价值的信息，这对工商银行客户服务的提升会有很大的帮助。

2. 传统客户服务分析流程

首先了解一下传统银行客户服务的分析流程。当客户拨打 95588 热线电话之后，客服座席会把他说的话和要求记录下来，存到客户之声系统之中，系统会对结构化的部分进行分析，比如投诉的数量、客户对我们满意度的打分或问题处理时效。

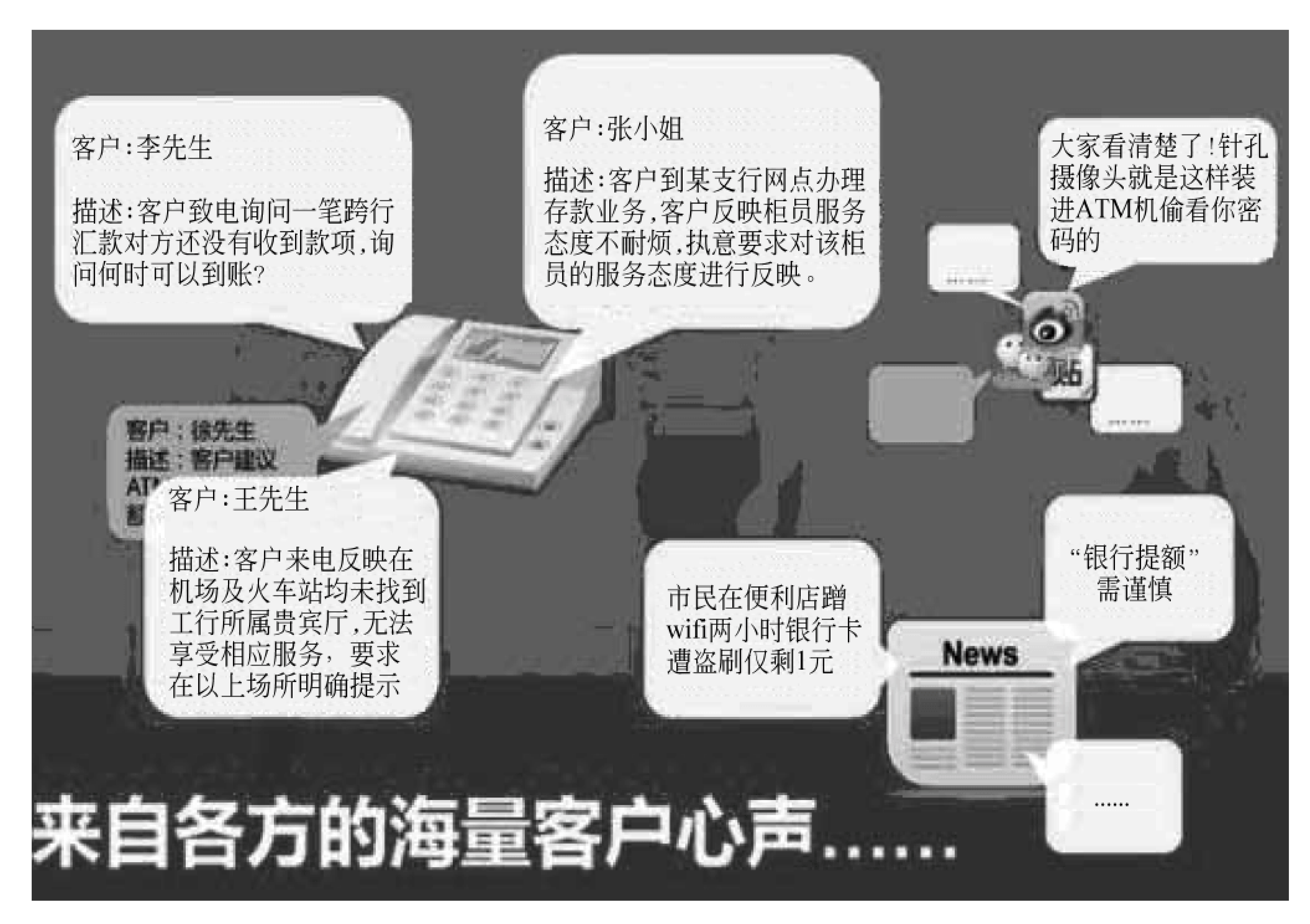


图 9.1 来自各方的海量的客户心声

对于其中非结构化数据的部分,就是客户说了什么当时没办法做自动分析,这只能由分析人员逐个来看,但毕竟数量比较多,人工阅读做不到非常全面,只能做抽查,大概看看客户在说什么。我们监测分析人员同时还会去登录一些新闻网站了解一下近期有没有与工行相关的事情发生,然后他会把这个情况记录下来,人工编写这么一个服务的报告。当时对我们的社交媒体是没有办法做到关注的。传统的银行客户服务分析流程如图 9.2 所示。

3. 结合文本挖掘的客户服务分析流程

在结合了文本挖掘技术之后有了一些流程变化,不仅对结构化数据做分析,同时也能够从客户反馈的文本当中提取出客户的热点意见,再把热点去和结构化数据做关联分析,就能得到更加丰富的分析场景。

同时,我们又新建了一套互联网的监测分析系统,能够对互联网上的金融网站和社交媒体网站做到自动的监控和分析,当然有些重要的事情发生的时候可以自动的形成监测报告。

从刚才服务流程的演变可以看到有了一些挖掘的功能,首先从技术来说丰富了分析的手段,原来只能对结构化进行分析,现在能够对文本数据客户所说的内容进行分析;其次扩大了分析的范围,原来只能关注到工商银行官方服务渠道所记录下来的信息,现在能够关注到在互联网上所传播的信息;最后是提升了分析的效率,原来需要员工逐条阅读工单,现在机器自动阅读。结合文本挖掘的客户服务分析流程如图 9.3 所示。

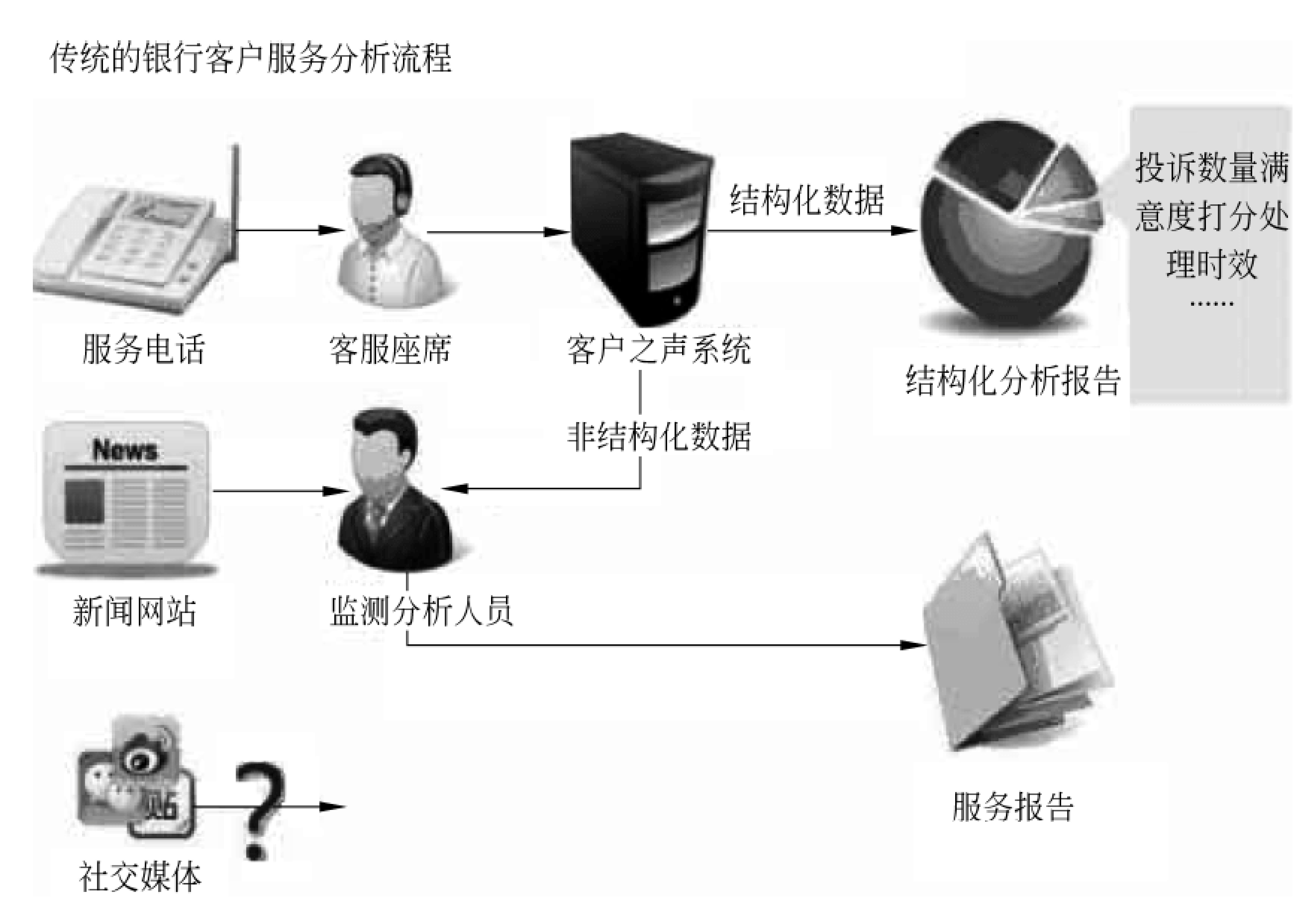


图 9.2　传统银行客户服务分析流程

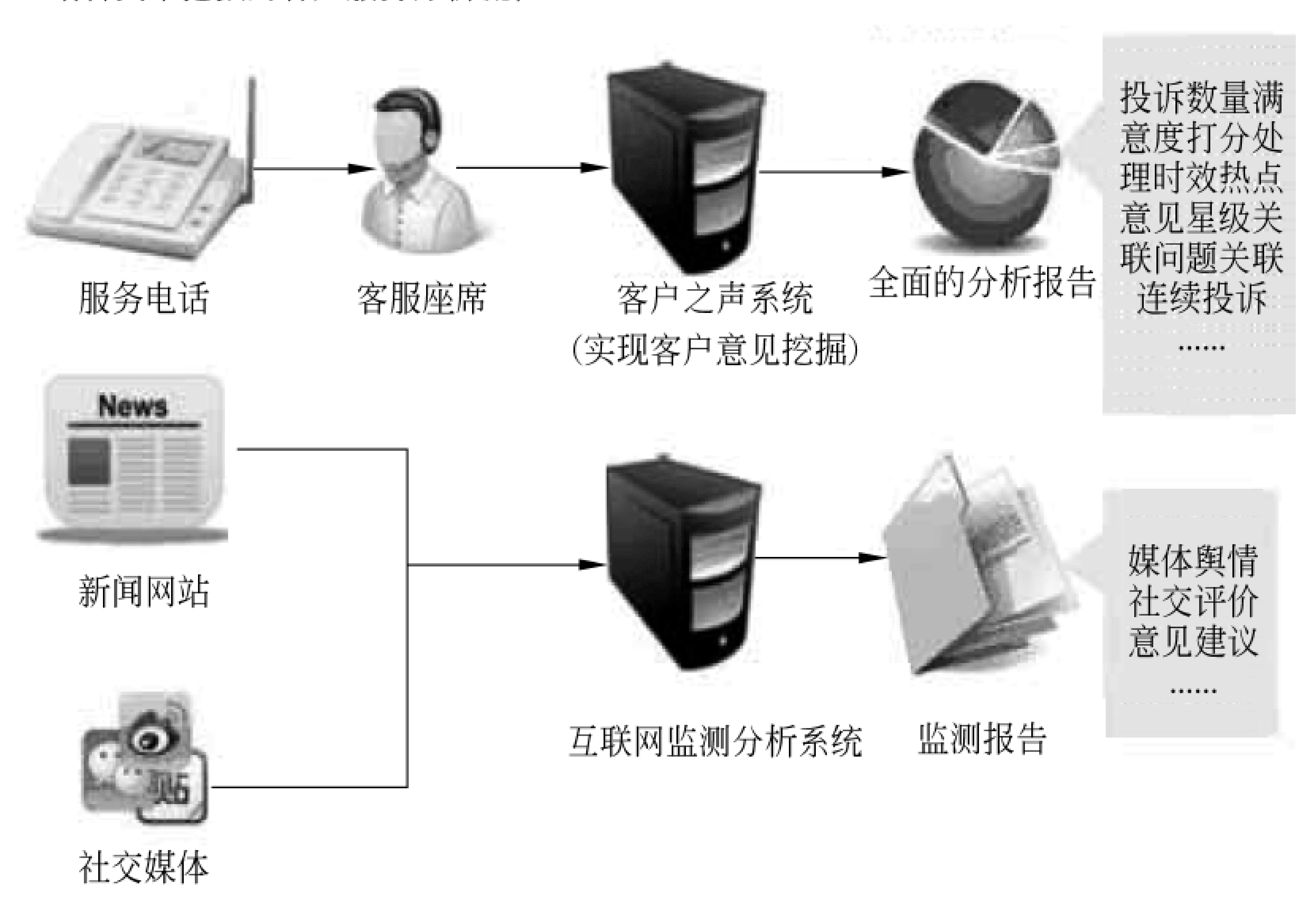


图 9.3　结合文本挖掘的客户服务分析流程

4. 客户意见挖掘业务价值

这些技术提升点之后就能在打响的文本反馈当中发现客户的热点意见集中在哪些方面，如果能够对这些客户所反映的共性问题主动发起一些措施，优化业务流程，可以提升客户满意度和客户忠诚度，而另一方面这些来电的投诉量会进一步的减少，也就从另一方面降低服务成本，减少了二次被动的服务投入。客户意见挖掘业务的价值分析如图 9.4 所示。

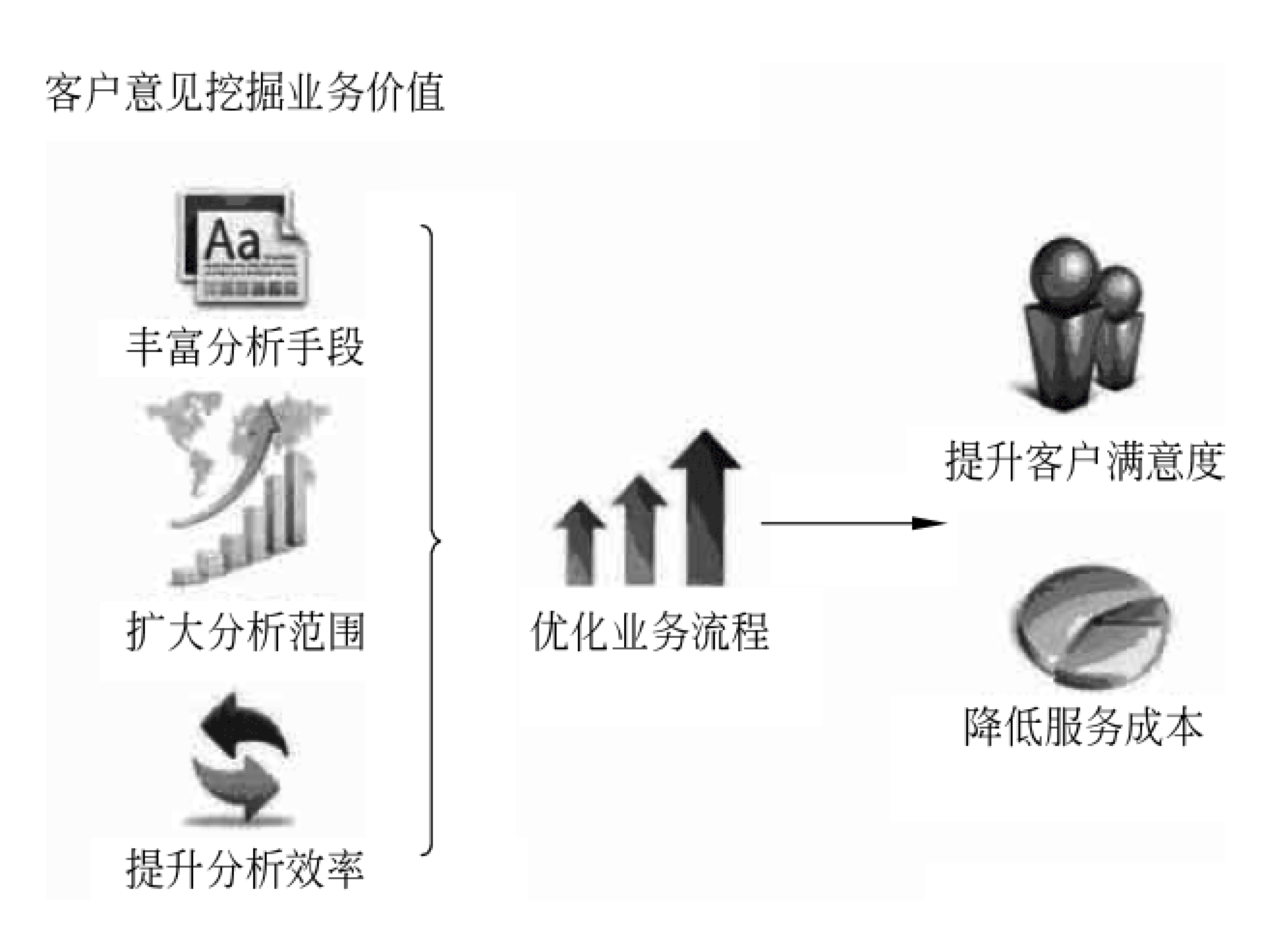


图 9.4　客户意见挖掘业务价值

9.1.3　银行风险管理

1. 从信用卡账单刷卡数据中,我们可以分析出什么

对于刷卡消费类的数据分析,如果能够拿到所有人的信用卡消费数据(一个人可能有多张信用卡),那么拿到这些信用卡消费数据应该如何展开分析,如图 9.5 所示。



图 9.5　银行信用卡风险管理

对于用户消费行为分析谈得比较多的思路仍然是需要首先搞清楚分析的目标,然后再根据目标的分析去采集和处理需要的数据信息。即数据分析本身是 KPI 驱动的,那么如果从最原始的数据明细入手,应该如何进行展开和数据维度的拓展?

对于有信用卡的人,我们收到的信用卡账单,往往有最简单的消费明细数据,如下:

消费清单(持有卡人卡号,姓名,消费商家,消费时间,消费金额)

可以看到这个消费明细数据本身是相对简单的,如果不结合其他的数据维度,单纯地去做统计分析并不会产生太多的作用。任何数据分析都需要结合对原始数据的维度拓展上,维度拓展后整个数据模型会更加丰富,则可以产生多维度的分析和数据聚合,如

图 9.6 所示。



图 9.6　信用卡账单刷卡数据分析

对于上面的消费详细清单数据，简单来看可以进行如下扩展：

人员信息（人员姓名，身份证号，年龄，姓名，职业类型，居住地址，家庭信息）

商家信息（商家名称，商家地址，商家经营类型）

有了人员信息就有第一层拓展，即对数据的聚合可以基于人员的属性维度，即我们拿到的消费明细数据，可以按照消费者性别、年龄段、职业类型等进行聚合。对于人员的识别唯一码不是姓名，而是人员的身份证号码，即通过身份证号码可以对一人多张信用卡的消费数据进行聚合。

有了商家信息，就可以根据商家的经营类型对不同类型的消费数据进行聚合。同时可以看到，对于商家详细地址信息本身是无法进行聚合的。那就要考虑在主体对象的属性中的单个属性本身的层次扩展，即地址信息可以进行扩展，即城市→区→区域→消费区域→商圈→大商场→具体地址。

如果地址有了这个扩展，就可以看到最终的消费数据可以做到按消费区域进行聚合，我们可以分析某一个商圈或商场的消费汇总数据，而这个数据本身则是从原始消费明细数据中进行模型扩展出来的。

可以看到，任何动态的消费明细数据，必须要配合大量的基础主数据，这些基础主数据可能有表格结构也可能是维度结构，这些数据必须要整理出来并关联映射上详细的消费明细数据。这样，最终的消费数据才容易进行多维度的分析。

消费时间本身也是重要的维度，可以根据时间段进行数据汇总，同时时间本身可以按年、按季度、按月逐层展开，也是一种可以层次化展开的结构。同时应注意到时间本身还可以进行消费频度的分析，即某一个时间段里面的刷卡次数数据，根据消费频度可以反推到某一个区域本身在某些时间段的热度信息。

如果仅仅是信用卡的刷卡消费清单数据，我们比较难以定位到具体的商品 SKU 信息上，如果是一个大型超市，则对于详细的用户消费购买数据，还可以明细到具体的商品

上，则商品本身的维度属性展开又是可以进行拓展分析和聚合的内容。

数据本身可能具备相关性，刷卡消费的数据往往可以和其他数据直接发生相关性，比如一个地区发出的大事件、在一个区域举办的营销活动以及从交通部门获取到的某个区域的交通流量数据。这些都可能和最终的消费数据发生某种意义上的相关性。

如果仅仅是从刷卡数据本身，前面谈到可以根据商户定位到商家的经营范围，究竟是餐饮类的还是服装类的。根据不同的经营类型可以分别统计刷卡消费数据，然后就可以分析，对于餐饮类的消费金额增加的时候服装类的消费是否会增加，即餐饮商家究竟对一个商场的其他用品的销售有无带动作用等？

同样的道理，对于人员可以分析不同年龄段的人员的消费数据之间是否存在一定的相关性，这些相关性究竟存在于哪些类型的商品销售上等。这些分析将方便我们制定更加有效的针对性营销策略。

2. 信用卡客户价值分析

让历史告诉未来。客户价值分析就是通过数学模型由客户历史数据预测客户未来购买力，这是数据挖掘与数据分析中一个重要的研究和应用方向。RMF 方法就是让历史告诉未来的趋势分析法，利用 RMF 方法科学地预测老客户未来的购买金额，然后对产品成本、关系营销费用等进行推算，即可按年、按季度、按月预测出客户未来价值。这里以信用卡为例，讨论和分析信用卡客户价值。

1）预测模型

对银行而言，预测客户未来价值能够使银行将传统的整体大众营销推进到分层差异化营销、一对一差异化营销的高度，对不同的分层客户采取不同的营销模式、产品策略和服务价格，从而推动和促进客户购买交易。

根据 RFM 方法，“客户价值”预测模型为：

客户未来价值＝银行未来收益－未来产品成本－未来关系营销费用

对于信用卡客户，我们定义此处的“未来”是指未来一年(也可以是未来一季度)。“银行收益”包括信用卡年费、商户佣金、逾期利息以及其他手续费等；“产品成本”即产品研发、维护和服务成本，包括发卡、制卡、换卡和邮寄等费用以及其他服务费用；“关系营销费用”即关系维护和营销成本，包括商户活动、积分礼品兑换、营销宣传等。

RFM 方法是目前国际上最成熟、最通用、最被接受的客户价值分析的主流预测方法。实际上，RFM 方法是一整套客户价值分析方法中的一部分(其中，R：最近购买日 Recency，F：购买频率 Frequency，M：平均单次购买金额 Monetary)，但是 RFM 方法最具有代表性，其他还包括客户购买行为随机过程模型、马可夫链状态移转矩阵方法、贝氏几率推导状态移转概率方法和拟合回归分析方法等。

(1) 预测未来收益。

由于“银行收益”包括信用卡年费、商户佣金、逾期利息以及其他手续费等，这里统一称为“购买金额”。因此，“客户未来购买金额”预测模型为：

$$\text{客户未来购买金额}=\text{未来购买频率}\times\text{未来平均金额}\times\text{未来购买频率概率}\times\text{未来平均金额概率}$$

其中，未来购买频率、未来平均金额、未来购买频率概率、未来平均金额概率均可通过客户

购买行为的随机过程模型来描述和求解。对于信用卡客户，“客户购买行为”包括刷卡、透支、取现、支付、分期等，以及客户消费习惯、还款习惯、收入贡献、信用额度、用卡来往区间、逾期时长、客户服务和副卡的客户购买行为等。

根据 RFM 方法预测过程，随机过程模型除了推导和计算客户未来购买频率概率、未来平均金额概率的密度分配之外，还隐藏着客户未来购买频率、未来平均金额的状态移转期望值和概率。因此，除了使用随机过程模型之外，还需使用贝氏几率方法推导状态移转期望值和概率。

此外，要科学地分析和预测客户未来价值，有必要用长度和宽度的二维样本数据建立一套牢固、可靠的随机过程模型，样本越大，客户未来价值的预测结果就越接近未来的事实。其中二维样本数据是指客户购买频率与购买金额是两个相互独立的不同的行为维度，不具有相关性。

(2) 预测未来产品成本和关系营销费用。

RFM 方法只能预测客户未来购买金额(或银行未来收益情况)，却不能预测出未来产品成本和关系营销费用。而采取平均法或移动平均法将客户历史价值、历史关系营销费用直接应用到客户未来，显然不适合；同样，采取 RFM 方法的概率分析方法来推断客户未来价值也是不适合的。因为未来产品成本和未来关系营销费用并不是源自客户的随机行为，而是由银行整体产品成本控制和差异化营销决定的，其未来变化不一定具有平滑趋势，未来客户的情况可能会出现逆反或抖动。因此，预测未来产品成本和关系营销费用需要采取其他方法。

首先要明确，未来产品成本和未来关系营销费用并不是随机现象，而是遵循各自发生的规律；且客户未来关系营销费用服从客户历史关系营销费用与购买金额的比例，即服从关系营销投入产出比。对于信用卡客户而言，通常以“年”为最小期数进行分析和预测，历史区间和未来区间是连续的，即两者之间无交易期数。所以，未来产品成本和未来关系营销费用的变化符合银行整体产品成本和营销费用的线性拟合回归规律。

因此，对于信用卡客户，“未来产品成本”预测模型为：

$$\text{未来产品成本}=\text{未来购买金额}\times(1-\text{CRM 毛利率})$$

$$\text{CRM 毛利}=\text{购买金额}-\text{产品成本}-\text{关系营销费用}$$

对于“未来关系营销费用”，定义：

$$\text{Ratei}=\sum\text{客户历史关系营销费用}\Big/\sum\text{客户历史购买金额}$$

$$\text{Expensei}=\text{客户历史最小关系营销费用(须大于 0)}$$

$$\text{Monetaryi}=\text{客户未来购买金额}$$

$$X=\text{Monetaryi}\times\text{Ratei}$$

因此，如果 $X>$Expensei，那么“未来关系营销费用”$=X$。否则，如果 Monetaryi$<$Expensei，那么“未来关系营销费用”$=X$；如果 Monetaryi$\geqslant$Expensei，那么“未来关系营销费用”$=$Expensei。

2) 客户价值

从以上分析可知，客户价值＝CRM 毛利＝购买金额－产品成本－关系营销费用。因此，在完整的客户关系生命周期内(即从建立关系到未流失的最近一次交易)，分析客户

未来价值的意义远远大于分析客户历史价值,因此通常意义上的客户价值分析就是对客户未来的价值进行分析和预测。

对于预测出的客户未来价值的结果,可按客户价值分层,并将传统的整体大众营销推进到分层差异化营销、一对一差异化营销的高度,其立足点就是客户价值的差异化分析。

通过分析和预测客户未来价值,即可清楚一旦高端客户、大客户流失将会造成未来怎样的利润损失,也可以挖掘出那些临近亏损或负价值的客户,并进行置疑分析,找出对策。但同时也要认识到,即使预测出客户的未来价值较高,也只能说明其价值势能(即潜在购买力)较高,坐等客户主动上门的价值动能(实际购买力)是不现实的,还需要通过其他沟通交流和营销渠道(如人工座席外呼、短信发送、微博私信、微信、邮件推送等)与客户互动,推动客户追加购买、交叉购买。

9.2 保险业应用

9.2.1 保险产业拥抱"大数据时代"或带来颠覆性变革

当今,数据已经渗透到每一个行业和业务领域,成为重要的生产因素。人们对于海量数据的挖掘和运用,预示着新一波生产率增长和消费者盈余浪潮的到来。中国的保险销售模式正在酝酿新的变革,互联网、大数据时代的到来给金融业造成的革命性、颠覆性的变化正在发酵,对保险业数据驾驭能力提出了新的挑战,也为保险业的大发展提供了前所未有的空间和潜力。

1. 深入挖掘大数据应用潜质

目前,大多数保险企业都已经认识到"大数据"改善决策流程和业务成效的潜能,但却不知道该如何入手,部分企业在"大数据"的时代浪潮下积极探索,成为先行者。2010 年,阳光保险集团建成数据挖掘系统,这在保险行业是第一家。利用该系统,开展了许多保险大数据智慧应用的项目,获得了一些成果,同时培养出了国内保险行业的第一批数据挖掘师。

通过深度挖掘和开发数据资源,提供可以用作产品定价的、承保口径的逐单数据,系统的行业终极赔付分析以及符合中国本土市场的财产险风险曲线,直保公司可以根据这些数据来分析某类风险的保险费率水平,了解公司与行业合理定价水平的差距,促进理性分析经营。同时,分析结果还可以应用到营销、业务拓展等方面,为直保公司决策提供参考。

2013 年,中国财险再保险公司行业数据分析中心正式挂牌成立,这是保险企业追赶"大数据"时代浪潮的一次标志性事件。早在 1996 年中再保险公司就利用与直保公司的非竞争关系,积极对数据进行集约化管理,拓展与直保公司在数据分析领域的合作。

大数据应用的关键是理念。思维转变了,数据就能被巧妙地用来激发新产品和新型服务。举一个利用与不利用数据结果相去甚远的例子:"淘宝现有一种运费保险,即淘宝买家退货时产生的退货运费原本由买家承担,如果买家购买了运费保险,退货运费由保险公司来承担。这种购买的结果是保险公司经营亏损很严重,直接导致它们不愿意再发展

和扩大运费保险。”运费保险真的必然亏损吗？答案是否定的。保险公司设计一套大数据智慧应用的解决方案：“因为退货发生的概率，跟买家的习惯、卖家的习惯、商品的品种、商品的价值、淘宝的促销活动等都有关系，所以，使用以上种种数据，应用数据挖掘的方法，建立退货发生的概率模型，植入系统就可以在每一笔交易发生的时候，给出不同的保险费率，使保险费的收取，与退货发生的概率相匹配，这样运费险就不会亏损了。在此基础上，保险公司才有可能通过运费险扩大客户覆盖面。”由严重亏损到成本控制得当并获取客户，靠的就是通过分析，挖掘大数据所提供的价值，吸引客户。

2. 大数据网络保险时代来临

大数据发展的障碍，在于数据的“流动性”和“可获取性”，而网络完美地解决了这个问题。通过网络对大数据进行收集、发布、分析、预测会使决策更为精准，释放更多数据的隐藏价值。与传统保险方式相比，网络保险具有降低保险公司和保险中介机构运营成本，拓展保险公司和保险中介机构业务范围，新型营销手段，有价值的交互式交流工具，提供较高水平的信息服务，为客户提供便捷工具，使客户享受个性化服务，降低保险公司风险，更有效地保护客户隐私以及虚拟化的交易方式等特性。

从产品设计角度来说，大数据时代下的网络保险能最大程度地满足不同客户的个性化需求，网络保险能优化客户的体验，“大数据”能根据客户需求设计出真正让客户满意的产品和服务，两者结合则完全是“以客户为中心”的。

从大数据时代的网络销售优势来看，一是大数据时代保险网销具有最广泛的客户群，有最大的发展潜力。二是互联网具有信息量大、传导速度快、透明度高的特点，交易双方信息更为对称。通过建立新型的“自动式”网络服务系统，保户足不出户就可以方便快捷地从保险公司的服务系统上获取公司背景到具体保险产品的详细情况，还可以自由地选择所需要的保险公司及险种，并进行对比，能获得低价、高效服务。三是节省费用，降低成本。通过网络出售保险或提供服务，保险公司只需支付低廉的网络服务费，从而降低房租、佣金、薪资、印刷费、交通费、通讯费等成本的支出。四是数据管理方面的天然优势。保险市场专业化的深入、经营水平的提高、服务品质的提升，都要建立在对数据尤其对客户消费数据的深入挖掘和分析的基础之上。

可见，大数据时代下的网络保险有利于推动营销体制改革。多年来，我国一直以保险代理人作为保险推销体系的主体重点发展，在寿险推销方面形成了以寿险营销员为主体的寿险营销体系。但是，目前这种体制还存在较为突出的问题。因客户缺乏与保险公司的直接交流，会导致营销人员为急于获取保单而一味夸大投保的益处，隐瞒不足之处，给保险公司带来极大的道德风险，为保险业的长远发展埋下隐患。而且，保险营销人员素质良莠不齐，又给保险公司带来极大的业务风险。此外，现有营销机制还存在效率低下的弊端。

因此，在大数据时代下发展网络保险，可以快速便捷地进行信息收集、发布，完美地实现大数法则的精致应用。为公众提供低成本、高效率的保险服务。

3. 网络保险需多项配套支持

一是财政支持。在推进保险公司的信息化进程中，政府可采取诸如信息技术方面的

投资部分抵消税收,税前可以预留部分资金用于信息技术改造等一系列措施,激励和推进大数据网络保险信息化进程。

二是培育网络保险集市。网络保险集市就是在网络上提供一个场所,使客户能在这里找到大量的保险公司,方便了解各个公司的基本信息或查询各个保险公司的某一险种的有关信息,并对该险种的优劣进行对比分析,选择最佳的公司进行投保。网络保险集市不仅会给客户带来方便,同时也会扩大保险公司的影响和业务量。因此,保险公司应在保监会和保险协会的组织下,全力支持并在网络保险集市上展示自己,进一步推动我国网络保险集市的发展。

三是建设大数据中心。大数据中心需要保监会和保险行业进行战略性的顶层设计。首先是与我国标准化数据管理中心进行合作,制定出保险业数据标准化的制度。其次是通过 5～10 年的时间逐步完成行业数据标准化建设。同时设计出非线性融合关系数据,并能进一步扩展的数据库。此外是设计柔性的框架和接口。通过以上步骤逐步完成我国保险业大数据中心的建设。

四是开发适合的险种。利用网络收集数据形成大数据,利用大数法则设计客户需求的产品,通过网络销售产品,并根据客户反馈进一步修正产品,实现开发与销售完美互动。

五是吸纳优秀人才和对已有员工在职教育。许多保险公司有一个规定,即无论是管理人员还是技术人员都必须完成一定的保险任务。似乎这条规定能为公司增加一点业务量,但是它无形之中会把一些优秀的保险管理人员和技术人员拒之于门外。大数据时代需要一流的管理人才和技术人才,必须破除这条不成文的规定。同时还应该重视对已有员工进行保险专业知识、外语知识和信息技术知识再教育,通过再教育提高公司员工综合素质。

六是责任与自由并举的信息管理。调查显示,66%的被调查者最关心投保后支付保费的转账安全性。消费者对于网络消费的顾虑心理主要集中在对网上交易安全和个人隐私保护的担忧上。因此,网络保险应格外注重网络安全,实现责任与自由的矛盾的和谐统一。

9.2.2 保险欺诈识别

没有核保压力,网销意外险领域更易出现欺诈案件;随着欺诈手法的复杂化,反欺诈也需用到大数据进行智能化反击。除了欺诈案件高发的车险领域,当前,保险欺诈正在向更大领域蔓延,在意外险、互联网保险以及农业保险等诸多领域,保险欺诈也正在显露苗头。

上海保监局面对保险欺诈,则充分利用上海保险业的信息平台技术优势,推行大数据智能化反保险欺诈工作模式,有效打击保险欺诈。

1. 网销意外险更易发生欺诈

深圳保监局日前发布消息称,该地区发生了几起互联网保险欺诈案件,在回复《证券日报》记者采访时,该局称,这几起案件的涉案金额不大,目前尚未结案,且未有法院判决结果,因此,目前尚不便公开具体案情。但业界人士认为,互联网保险欺诈风险事实上已

经显露出苗头。

此前，安徽保监局也发布消息称，互联网保险存在较大的道德风险，“短意险客户可以通过网络购买不同公司的短期意外险产品，目前已查明个别高风险客户在多个公司累计投保金额超过千万元”，安徽保监局在其调研报告中特别强调了网销短期意外险发生的风险，数据也表明，意外险正是目前网销保险最主要的品种。

江苏省公布的 2007—2013 年江苏十大典型保险欺诈案件中，其中一件即为投保人以本人作为被保险人，投保了 1000 余万元的人身意外伤害保险。此后，该投保人买来排骨，在剁排骨时故意将自己左手食指近节指端剁断，被鉴定为七级伤残。在到保险公司索赔过程中案发，该投保人被法院以保险欺诈罪(未遂)判处有期徒刑 6 年，并处罚金 5 万元。

业内人士认为，实际上，很多保险公司无论是网销意外险还是其他渠道销售该险种，都没有严格的核保流程，也缺乏相应的技术手段了解投保人在其他平台的投保情况，因此，意外险的欺诈风险并非网销渠道专属，不过，对于投保人而言，通过网络购买比其他方式更加方便，也缺少被核保的心理障碍，因此更容易通过这种方式实施保险欺诈。

2. 大数据智能反欺诈兴起

对于保险公司和监管层而言，一方面要解决理赔难问题，另一方面也须遏制保险欺诈。为净化保险环境，遏制保险欺诈，不少地方采取了多种措施，包括不同部门联手打击保险欺诈，建设信息平台杜绝信息孤岛等方式。

“保险公司内控薄弱，是保险欺诈案件时有发生的主要原因。”吉林保监局在调研报告中指出。为此，各地在反保险欺诈工作中，不仅要求保险公司加强内控，同时，针对保险欺诈涉及人员多等特点，反保险欺诈工作还通常与公安、法院、检察院等部门形成常态合作机制，例如，陕西就建立并完善了“高风险修理厂数据库”“高风险客户数据库”和“高风险从业人员数据库”，为保险公司提供预警和服务。陕西反保险欺诈中心成立一年以来，各公司共向中心送报可疑线索 1819 件，涉及金额约 7000 万元；全省公安经侦部门侦破保险欺诈件 29 起，涉案金额 603 万元。因涉嫌保险欺诈，投保人或被保险人主动放弃索赔或公司拒赔案件达 1368 件，为保险公司预防和挽回经济损失达 5531 万元。

上海保监局面对保险欺诈，则充分利用上海保险业的信息平台技术优势，依托“机动车辆保险联合信息平台”、“人身险综合信息平台”和“道路交通事故检验鉴定信息系统”，推行大数据智能化反保险欺诈工作模式，具体包括利用大数据方式进行风险预警、关联排查以及数据串并，通过这些方式有效打击保险欺诈。近期，上海保险行业识别并移送了一起勾结二手车商贩、故意制造交通事故的“一条龙”车险团伙欺诈案，经线索串并排查后，案件涉及赔案 60 余起、总金额超过 100 万元，涉及人员 20 余人。该案经公安机关侦破，目前主要犯罪嫌疑人已被判刑。

针对互联网时代对的保险监管，江西保监局提出，传统保险监管无法完全满足互联网保险的监管要求，互联网保险带来的新风险需要专业的风险监测和管控。虚拟网络世界跨省跨地域，需要进一步整合监管资源，保险监管在属地化管理过程中面临跨省监管的问题，需要上下联动、各局配合，委托监管和检查。该局还指出，要加强保险、银行和证券的监管合理，提升监管效能，还需要建立保险业网络征信数据库反保险欺诈。

9.3 证券期货应用

9.3.1 安徽使用大数据监管证券期货

多年以来，股市的波动牵动着大家的心，目前安徽正在使用大数据“电子眼”对证券期货市场进行监管，60%的违规运作都是通过大数据抓取发现的。

1. 违规证券期货难逃“电子眼”

普通商家的监管有工商的例行检查，但是金融领域的违规不免有一些隐蔽性。然而，这样的隐蔽性如果未被发现，会给投资者带来巨大的经济损失。安徽正在用科技手段破解这一难题。

“证券期货公司的各项数据都在系统之中，如果有人试图进行违规操作，大数据都会发现。这个大数据系统的违规抓取成功率非常高，‘电子眼’并不那么好骗。目前我们查处的违规行为中，60%都是来自大数据的发现。”

2015 年，安徽省实施稽查提前介入 1 家次，将 4 起违法违规线索移送稽查。对市场主体采取行政监管措施更是达到了 9 次；开展案件调查 22 件(含辖区案件 14 件)，同比增长 10%，包括证监会法网专项行动 A 类案件 4 件、涉外案件 1 件、与公安部门共同查办案件 1 件、移送公安部门案件 1 件。

而对于日渐火热的贵金属市场，安徽也在大力监管，去年共处理非法贵金属交易等违法违规证券期货活动线索 12 起，其中，5 起移送公安或工商部门；依法审理 5 起行政处罚案件，执行罚没款 140 万元。

在去年进行的 47 家次现场检查中，安徽的范围也在扩展。检查对象不仅包括证券期货经营机构、投资咨询机构，也包括了风生水起但又风险重重的互联网股权众筹融资平台，对 1 名从业人员、2 家机构采取了行政监管措施。

在股市频繁波动的背景之下，安徽针对违规减持行为，及时采取监管措施，并移送稽查部门查处。同时，加强对证券期货经营机构信用业务、资管业务等核心业务，以及上市公司大股东股权质押风险的监控。目前，安徽辖区 60 家公司制定了维护股价稳定方案，31 家公司实施了增持，增持总额约 13 亿元。

2. 融资“倒金字塔”结构正在改变

对于企业融资问题，不同规模的企业或许有着不同的选择。但是，目前安徽融资渠道不断拓宽，过去的“倒金字塔”结构正在发生着变化。

2015 年，安徽省新增 IPO 辅导备案企业 42 家、申报企业 9 家，8 家公司成功上市，全省境内上市公司达到 88 家，家数超过湖北，跃居中部第一、全国第九；新三板挂牌企业新增 117 家，达到 162 家，家数居全国第九；省股权托管交易中心新增挂牌企业 488 家、托管企业 611 家，总数分别达到 710 家和 861 家。

2015 年，全省资本市场完成直接融资 358.32 亿元，同比增长 60%，融资额与我省历史最高水平基本持平(2011 年融资 358.36 亿元)。其中，8 家公司 IPO，融资 40.12 亿元；

13 家上市公司非公开发行股份，融资 178.21 亿元；59 家新三板挂牌公司开展 84 次定向发行，融资 20.41 亿元；10 家企业发行公司债券，融资 91 亿元；5 项资产支持证券成功发行，融资 28.58 亿元。

9.3.2　"大数据"分析挖出基金"老鼠仓"的启示

随着基金"老鼠仓"不断被揪出，"大数据"监管这个字眼也逐渐被投资者所熟悉。靠"大数据"这个利器，监管机构对内幕交易的稽查力度越来越大，今年以来基金经理变更数量和比例也明显高过往年。

将"大数据"分析挖掘应用到证券基金监管中绝对是方向，绝对是远远超越传统监管方式的一把高科技监管利器。对于监管部门利用"大数据"利器，在挖出基金老鼠仓上小试牛刀却大获全胜的做法给予充分肯定。

证券期货基金市场无论是投资者开户，还是交易；无论是交易场所，还是投资分析；无论是股票期货基金托管，还是交易资金银行第三方存管，所有交易活动完全是网络电子化的。任何投资者只要发生交易活动，都会在网络上留下足迹，并且，这种足迹可以追查寻觅到每一个具体的投资者"本人"。这就为"大数据"在资本市场的任何运用奠定了基础，"大数据"可以在资本市场发挥几乎是无所不能的作用，包括挖出基金老鼠仓。

这与传统监管手段完全处于被动地位相比较，简直是一个质的变化和大飞跃。主要区别是传统监管方式是被动的，效率极低，隔墙扔砖头、砸着谁是谁，逮住的是个别的、是虾米，放走的是大多数、是大鱼。传统监管方式主要有两种：一是监管部门人员突然袭击，出现在证券基金公司，让所有人员立即离开，然后在证券基金工作人员电脑中现场检查发现线索。二是依靠内部举报。其中有些老鼠仓是基金经理的"小三"举报，还有配偶因为离婚财产分配不均举报、办公室斗争的同事举报等。

"大数据"分析挖出基金老鼠仓，监管方式是主动的、全面的、高效的，不会放过任何一个老鼠仓。"大数据"用来挖老鼠仓，主要是基于沪深两大交易所每天的海量数据，根据老鼠仓的主要特征，筛选出若干种最具老鼠仓特征的数据指标，在沪深两大交易所海量数据平台上无时无刻进行抓取。正如"大数据"的鼻祖美国第二大百货公司——塔吉特为了获取孕妇信息而最早投放广告抢夺孕妇客户一样。根据怀孕者的消费习惯筛选了 20 多种产品，通过"大数据"抓取分析，最终截获客户，获得成功。从现有的公开资料来看，监管机构的"大数据"主要是沪深两大交易所各自掌握的监测系统，主要分为对内部交易的监察、对重大事项交易的监察、联动监察机制和实时监察机制四个方面。这套监控系统有着所谓的"大数据"分析能力，并有实时报警等功能，主要是对盘中的异常表现进行跟踪和判断。

这是传统抓老鼠仓方式不可比拟的。传统监管方式就像用一个鱼钩垂钓一样，是被动地坐等鱼儿上钩。而借助"大数据"挖掘分析监管方式，就像向大海中撒了一张大网，一旦有异常情况就可以自动收网。

监管部门必须转变监管思路。过去那种运动式、集中行动式的人海战术监管方式，必须转变为互联网思维、互联网金融、大数据模式的高效主动监管方式。有报道说，证监会正在扩大稽查总队的阵容，人数或将在 300 人的基础上再扩编 300 人。动不动就增加人

员、采取人海战术的做法还是传统思维在作怪。阿里小贷完全借助于大数据挖掘，只有300多个员工，就给70万家小微企业放贷款，累计放贷已经超过1000多亿元。这是传统银行不可想象的。拼的是高效高科技手段的大数据，而不是人海战术。

随着互联网的普及特别是移动互联网的迅猛发展，所有社会经济文化等活动都将互联网化，都将由线下搬上网络。这就意味着无论是自然人、社会人还是法人的所有足迹都将广泛、越来越多地在网络上留下印记和足迹。通过“大数据”对这些足迹进行挖掘，将会挖出一座大金矿。

“大数据”挖出最大老鼠仓启示我们，“大数据”不仅具有商业挖掘价值，而且也是监管经济金融活动甚至是反腐败的利器。官员及其家属亲朋好友的通信、经济活动、财富存款、消费社会足迹等都可以通过“大数据”挖掘出来。比如，银行、证券、基金等系统已经比较完善，未来不动产也将全国联网，这就将使得官员以及家属亲朋好友的一切家庭个人财务活动都将在网络上通过大数据分析可以挖掘出来，一旦发生异常，就将成为发现腐败的重要线索。

总之，“大数据”挖出最大老鼠仓启示我们，“大数据”应该尽快上升到国家战略，作为重大科技项目全力推进。不仅仅是为了“大数据”科技和经济，也是反腐败的利器，具有重要的政治价值。

9.4 金融行业应用

9.4.1 汽车金融公司怎么实现大数据管理

1. 汽车金融与大数据的关系

在谈汽车金融与大数据的关系前，觉得有必要对汽车金融进行一个“菜鸟”解读，汽车好懂，但是对于金融的理解，可以用三句话做最好的解读：

(1) 为有钱人理财，为缺钱人融资(金融是有资金流动的行为)；

(2) “信用”“杠杆”“风险”(三者缺一不可，相辅相成)；

(3) 金融如果不为实体经济服务，就是毫无意义的泡沫(金融依托于实体经济，也可能产生泡沫)。

无论是何种金融行为，均是建立在信用基础上的杠杆收益与风险的均衡，所以金融存在的基础是信用，汽车金融也不例外，就我国目前的征信法制建设水平而言，对客户的信用评价越准确，越容易把握杠杆与风险的均衡，就如同自己借十万给别人，对别人的收入、背景越了解，同等收益的条件下对风险的判断更有把握，那么对于借款人的信息了解程度就变得异常关键。

然而个人的信用评估和实现气象预测有非常类似之处，一个人或者群体的信用好坏取决于很多的变量，而且信用本身不是静态的，而是一个动态的行为特征的体现——资产、收入、消费、个性、习惯、社交网络等等都会对信用产生影响。在汽车金融行业，由于面对数量庞大的客户群体，如何从大量、多样、快速变化、低价值密度的信息中通过大数据对个体大量信用行为进行收集、整理、分析，把这些糅合在一起时，使得人的信用立体化，从

而甄别出价值客户，设计不同的金融产品获取最大化的收益，就变成了汽车金融行业发展的首要任务。

2. 汽车金融公司怎么实现大数据管理

目前个人汽车消费贷款方式有银行、汽车金融公司、整车厂财务公司、信用卡分期购车和汽车融资租赁五种，由于操作主体不同，对信息的需求以及积累的基础不一致，各个单位在实现大数据管理过程中可能存在各种差异，传统的金融业者由长期系统的金融服务积累的数据完全可以在确保用户隐私和商业机密的前提下，与各行各业通过数据间的共享、交换和买卖以生成大数据，在此之上探索全新的产品和服务，而对于一些依托于厂家或存在行业局限的汽车金融公司而言，如何规划并设计大数据管理之路，目前业内并无成熟经验可以借鉴，以下仅为工作思考中形成的几个观点，聊以抛砖引玉。

1）“外部大数据＋企业大数据”是必经之路，存量客户的信息挖掘是一笔宝藏

目前行业应用的最多的是人行的征信数据，而随着这几年的发展，市面上已经出现了各种信息咨询公司，有的可以提供工商数据，有的可以提供银联消费数据，有的可以提供小额贷款信用数据，所有这些，都是对以人行为中心的外部征信大数据的补充，这部分数据相对稳定，获取的渠道也比较透明，然而对于部分行业诸如商用车行业而言，由于客户购买车辆属于生产资料，行业经验的积累对于客户的盈利水平有明显的影响效果，同时生产资料都有淘汰更新的自然生命周期，而动辄几万，十几万的客户数量，已经把相应行业的朋友圈客户基本固定。这样一来，对于如何发掘重复购买的客户，实现价值营销具有重要的意义。

2）获取动态信息的渠道很关键，硬件与软件都重要

为了实时获取客户信息，防范客户风险，部分厂商在出厂的时候就预装了 GPS 设备，除去基本定位导航功能外，越来越多的厂商开始拓展其他功能，比如回传里程数、油耗、工况，实现远程诊断等等，逐步开始搭建自己品牌的车联网平台，但是就目前的实际情况来看，就算应用得比较好的商用车领域，各大厂商的软件系统、硬件设施等仍然存在数据质量差、回传效率低、不防拆等缺陷，出现各大车联网规划都很丰满，但是现实却很骨感的尴尬局面。

当然，行业内也有一些很具有前瞻性的公司，已经悄悄地开展了基于特定行业平台的大数据平台建设工作，通过平台的作用，整合上下游资源，把与汽车相关的保养、维修、换件、加油、保险甚至餐饮整合至特定平台，通过以人为中心的数据库建立，稳定客户资源，并根据客户在平台上的大数据条件，为客户提供金融贷款等服务。

3）大数据建设可以从特定行业或特定区域开始，再实现跨行业跨区域的大数据整合

前面说过，大数据具有低价值密度的特性，在大数据建设的初期，应对数据的使用维度进行规划，重点收集哪几个维度的信息，并把信息根据性质划分重要程度，实现外部信息和内部信息相结合，必备信息和补充信息相交错，静态信息和动态信息相辅相成，然而中国本来就是一个人口大国，如若全面铺开，相信暂时没有哪个企业能做到大而全的信息收集，但是在特定区域、特定行业，客户的相关信息相对固定，收集渠道相对稳定，这样对于特定领域的数据信息整合提供了可操作性，一旦细分行业细分客户的大数据成型，随着

规模扩大，即可以与相关行业实现数据共享或交换，在信息得到不断挖掘之后，可以想象，在未来的某一天，行业内可能出现1到2家具备垄断性质的大数据平台。

4）信息安全的敏感性对大数据发展提出更规范化的管理要求

2013年"棱镜门"事件暴露了美国情报机关正在利用大数据技术，对全球通信系统和互联网实行全面的实时监控，进行大数据采集、挖掘、分析、关联，引发了世界信息安全危机。在全球规模庞大的信息泄露关联产业，一批黑客长期从事截获并贩卖大众信息的工作，而接货者则通过计算机自动比对，将买来的账号密码等信息在各大金融机构网站、电商网站进行"撞库"，成功率通常可以达到5%～10%，成功"撞库"的信息将高价卖出，以便下个团队用以挪走消费者资金、非法支付和欺诈勒索等等。大数据时代如何兼顾安全与自由、商业利益与个人隐私，从而推动科技的进步，实现可持续发展，是每个人都应该关心的话题。

9.4.2 大数据决定互联网金融未来

互联网金融不是互联网和金融的简单叠加，更深层次的变化是：一些基于互联网应用的特有技术，推动了新的商业模式、产品、服务、功能在金融业内出现，金融体系随之经历着新的变革。大数据就是其中的典型代表，它也被视为推动互联网金融发展的重要驱动力之一。

麦肯锡全球研究院在其发布的《大数据：创新、竞争和生产力的下一个新领域》报告中指出："大数据之'大'通常是指数据量大到超过传统数据处理工具的处理能力，是相对和动态的概念。此外，大数据又被引申为解决问题的方法，即通过收集、分析海量数据获得有价值信息，并通过实验、算法和模型，从而发现规律、收集有价值的见解和帮助形成新的商业模式。"

金融业是大数据的重要产生者，交易、报价、业绩报告、消费者研究报告、官方统计数据公报、调查、新闻报道无一不是数据来源。但反过来，大数据对于互联网金融发展的助推作用也逐渐浮现。

1. 目标用户拼精准

大数据对于互联网金融的第一个助推作用在于寻找合适的目标用户，实现精准营销。

互联网金融领域的新创企业或做贷款，或卖产品，凭借高额收益率、手续费优惠，吸引用户选择自己。然而，在越来越多同类企业吹响混战号角的同时，互联网金融企业也不得不面对来自同行业的竞争。盲目扩张，产品单一，使得竞争力不强的互联网金融企业，由于不能保证稳定流量、无法留住客户而倒闭，成为行业的"炮灰"。上海永利宝金融信息服务有限公司CEO余刚分享了一组数据，以互联网金融领域的P2P业务为例，截止到2013年底，中国有450家P2P公司，最短命的P2P企业出现在海南省，创立2天即倒闭。

在巨大市场压力面前，许多互联网金融企业都已意识到自身产品的营销策略很大程度上影响了企业的生存与发展。欲在竞争激烈的市场中占有一席之地，互联网金融企业需要更精准地定位产品，并推送给目标人群。正如德邦证券董事长姚文平在其《互联网金融》一书中指出的："与其一味地苦思如何'做得更好'，不如考虑如何'做得不同'"。

谁是潜在的购买者？如何找到他们？并让他们产生兴趣？

精准营销的实现程度是互联网金融企业存活与崛起的关键所在，这个领域虽然未达到成熟的发展状态，但确实已经有了一些有参考价值的营销案例。例如，梧桐理财网推出了2万元起点的“梧桐宝”，是一款8%～10%预期收益的互联网理财产品，其目标客户是能够承担“两万元起投”的中产阶级；速溶网推出的“速溶360”旨在为在校大学生及毕业生提供金融服务……

大数据在为这些互联网金融企业找到自己的目标客户，并解决精准营销的问题上发挥了重要作用。大数据通过动态定向技术查看互联网用户近期浏览过的理财网站，搜索过的关键词，通过浏览数据建立用户模型，进行产品实时推荐的优化投放，直击用户所需。

2. “芝麻信用”控风险

其次，大数据在加强风险可控性，支持精细化管理方面助推了互联网金融，尤其是信贷服务的发展。

通过分析大量的网络交易及行为数据，可对用户进行信用评估，这些信用评估可以帮助互联网金融企业对用户的还款意愿及还款能力做出结论，继而为用户提供快速授信及现金分期服务。

事实上一个人或一个群体的信用好坏取决于诸多变量，如收入、资产、个性、习惯等，且呈动态变化状态。可以说数据在个人信用体系中体现为“芝麻信用”，它便于解决陌生人之间以及商业交易场景中最基本的身份可信性问题，以及帮助互联网金融产品和服务的提供者识别风险与危机。这些数据广泛来源于网上银行、电商网站、社交网络、招聘网、婚介网、公积金社保网站、交通运输网站、搜索引擎，最终聚合形成个人身份认证、工作及教育背景认证、软信息（包括消费习惯、兴趣爱好、影响力、社交网络）等维度的信息。

支付宝的大数据服务部负责人李颖赟以支付宝的用户数据举例，目前支付宝3亿名实名认证用户覆盖了近一半的中国网民，他们的上网足迹提供了涵盖购物、支付、投资、生活、公益等上百种场景数据，每天产生的数据相当于5000个国家图书馆的信息量。当人们在淘宝、天猫等电子商务平台上进行消费时就会留下自己的信用数据，当这些信息积累到一定程度，再结合交易平台上用户的个人信息、口碑评价等进行量化处理后，就能形成用户的行为轨迹，这对还原每一个人的信用有相当大的作用。同时，通过交叉检验技术，辅以第三方确认客户信息的真实性，以及开发网络人际爬虫系统，突破地理距离的限制，可以更全面、更客观地得到风险评估结论，从而加强互联网金融服务风险的可审性与管理力度。

毫无疑问，大数据将在互联网金融将大展身手，但大数据只是分析工具，是人类设计的产物，不应过分迷信。以P2P借贷行业为例，目前借贷业务不仅需要网络审核，更需要线下审核，信贷员的从业经验和责任心是信贷安全的重要保障。另外，除了个别企业，大部分互联网金融企业目前的用户规模和交易额都不大，缺乏大数据基础，也无力承担所需的基础设施和处理成本。在互联网金融的发展过程中，如何发挥大数据的优势，避免其劣势，将决定互联网金融的未来。

3. 六种可用于互联网金融风险控制（征信）的大数据来源

近年来，以第三方支付、P2P平台、众筹为代表的互联网金融模式引起了人们的广泛

关注，该模式大量运用了搜索引擎、大数据、社交网络和云计算等技术，有效降低了市场信息不对称程度，大幅节省了信息处理的成本，让支付结算变得更便捷，达到了同资本市场直接融资、银行间接融资一样高的资源配置效率。但由于我国互联网金融出现的时间短，发展快，目前还没有形成完善的监控机制和信用体系，一旦现有互联网金融体系失控，将存在着巨大的风险。

首先是信用风险大。目前我国信用体系尚不完善，互联网金融的相关法律还有待配套，互联网金融违约成本较低，容易诱发恶意骗贷、卷款跑路等风险问题。特别是 P2P 网贷平台由于准入门槛低和缺乏监管，成为不法分子从事非法集资和诈骗等犯罪活动的温床。

其次是网络安全风险大。我国互联网安全问题突出，网络金融犯罪问题不容忽视。一旦遭遇黑客攻击，互联网金融的正常运作会受到影响。

互联网金融企业通过获得多渠道的大数据原料，利用数学运算和统计学的模型进行分析，从而评估出借款者的信用风险，典型的企业是美国的 Zest Finance。其通过分析模型对每位信贷申请人的上万条原始信息数据进行分析，并得出超过数万个可对其行为做出测量的指标，而这一过程在 5 秒钟内就能全部完成。在进行数据处理之前，对业务的理解、对数据的理解非常重要，这决定了要选取哪些数据原料进行数据挖掘，进入“数据工厂”之前的工作量通常要占到整个过程的 60%以上。

目前，可被用于助力互联网金融风险控制的数据存在多个来源。

一是电商大数据，以阿里巴巴为例，它已利用电商大数据建立了相对完善的风控数据挖掘系统，并通过旗下阿里巴巴、淘宝、天猫、支付宝等积累的大量交易数据作为基本原料，将数值输入网络行为评分模型，进行信用评级。

二是信用卡类大数据，此类大数据以信用卡申请年份、通过与否、授信额度、卡片种类、还款金额等都作为信用评级的参考数据。国内典型企业是成立于 2005 年的“我爱卡”，它利用自身积累的数据和流量优势，结合国外引入的 FICO（费埃哲）风控模型，从事互联网金融小额信贷业务。

三是社交网站大数据，典型企业为美国的 Lending Club，它基于社交平台上的应用搭建借贷双方平台，并利用社交网络关系数据和朋友之间的相互信任聚合人气，平台上的借款人被分为若干信用等级，但是却不必公布自己的信用历史。

四是小额贷款类大数据，目前可以充分利用的小贷风控数据包括信贷额度、违约记录等。由于单一企业信贷数据的数量级较低、地域性较强，业内共享数据的模式已正逐步被认可。

五是第三方支付大数据，支付是互联网金融行业的资金入口和结算通道，此类平台可基于用户消费数据做信用分析，支付方向、月支付额度、消费品牌都可以作为信用评级数据。

六是生活服务类网站大数据，包括水、电、煤气、物业费交纳等，此类数据客观真实地反映了个人基本信息，是信用评级中一种重要的数据类型。

9.4.3 移动大数据在互联网金融反欺诈领域的应用

根据《2015中国移动互联网发展指数报告》，中国共拥有12.4亿台移动端设备，其中移动智能手机的保有量为9亿，每个移动互联网用户拥有大概1.35部智能手机。移动互联网用户中80后、90后、00后占比超过了72%，成为移动互联网主要用户。平均每部手机装载了41款应用，平均每天打开25款应用，相对去年有较大的提升。

移动互联网正在影响着人们的生活，移动设备端产生的数据也蕴藏着巨大的商业价值。2014年美国移动设备位置信息产生的市场价值大概为1000亿美元，2015年中国移动大数据的市场刚刚开始。

1. 移动大数据的商业价值

在PC互联网时代，不管用户是否喜欢BAT，其网站仍然在那里。但是在移动互联网时代，如果一个用户不喜欢这个应用，就可以在2秒钟内删掉这个APP，彻底中断和它的连接，无论其是不是BAT。在移动互联网时代，选择权完转向用户，消费者将成为数字世界的中心。过去以品牌为中心的消费形式，将会转变为以消费者为中心的消费形式。

智能手机上安装的APP和APP使用的频率，可以代表用户的喜好。例如喜欢理财的客户，其智能手机上一定会安装理财APP，并经常使用；母婴人群也会安装和母婴相关的APP，频繁使用；商旅人群使用商旅APP的频率一定会高于其他移动用户。未来80后、90后将成为社会的主要消费人群，他们的消费行为将会以移动互联网为主，APP的安装和活跃数据更加能够反映出年轻人的消费偏好。

智能手机设备的位置信息代表了消费者的位置轨迹，通过这个轨迹可以推测出消费者的消费偏好和习惯。在美国，移动设备位置信息的商业化较为成熟，GPS数据正在帮助很多企业进行数据变现，提高社会运营效率。在中国，移动大数据的商业应用刚刚开始，并且在房地产业、零售行业、金融行业、市场分析等领域取得了一些成果。

特别在互联网金融领域的应用，移动大数据正在帮助互联网金融企业实施反欺诈，降低恶意诈骗给互联网金融企业带来的损失。

2. 恶意欺诈成为互联网金融的主要风险

近几年，互联网金融爆发式发展，预计2015年P2P的交易总额将会超过1万亿，将成为具有影响力的产业。最近半年，大量的金融行业专业人士和传统产业资本进入到互联网金融领域，表明这个产业的生命力正在不断增强，有的P2P企业的年交易额已经突破百亿元，有的P2P企业估值也超过了15亿美元。

但是在P2P行业，其面对的风险也在加大，除了传统的信用风险，其外部欺诈风险正在成为一个主要风险。有的P2P公司统计过，带给P2P公司的最大外部风险不是借款人的坏账，而是犯罪集团的恶意欺诈。网络犯罪正在成为P2P公司面临的主要威胁之一，甚至在一些P2P公司，恶意欺诈产生的损失占整体坏账的60%。很多P2P公司将主要精力放在如何预防恶意方面。高风险客户识别和黑名单成为预防恶意欺诈的主要手段。

3. 移动大数据在反欺诈领域的应用

移动大数据中的位置信息代表了用户轨迹，商业应用较早。2014年，美国移动设备位置信息的市场规模接近1000亿美金。但中国移动设备位置信息的商业应用才刚刚开始。

从技术上讲，定位移动设备的位置有三种方式：第一种是通过运营商的三个基站定位，其误差大概在200m；第二种是通过手机APP中的GPS位置信息定位，大概误差为50m；第三种是通过WiFi定位，误差大概在3～5m。在移动设备位置信息商业应用中，三种定位方式都被应用，室内以WiFi定位为主，室外以GPS定位为主。移动大数据在反欺诈领域具有以下应用场景。

1）用户居住地的辨别

线上的欺诈行为具有较高的隐蔽性，很难识别和侦测。P2P贷款用户很大一部分来源于线上，因此恶意欺诈事件发生在线上的风险远远大于线下。中国的很多数据处于封闭状态，P2P公司在客户真实信息验证方面面临较大的挑战。

移动大数据可以验证P2P客户的居住地点，例如某个客户在利用手机申请贷款时，填写自己居住地是上海。但是P2P企业依据其提供的手机设备信息，发现其过去三个月从来没有居住在上海，这个人提交的信息可能是假信息，发生恶意欺诈的风险较高。

移动设备的位置信息可以辨识出设备持有人的居住地点，帮助P2P公司验证贷款申请人的居住地。

2）用户工作地点的验证

借款用户的工作单位是用户还款能力的强相关信息，具有高薪工作的用户，其贷款信用违约率较低。这些客户成为很多贷款平台积极争取的客户，也是恶意欺诈团伙主要假冒的客户。

某个用户在申请贷款时，如果声明自己是工作在上海陆家嘴金融企业的高薪人士，其贷款审批会很快并且额度也会较高。但是P2P公司利用移动大数据，发现这个用户在过去的三个月里面，从来没有出现在陆家嘴，大多数时间在城乡接合处活动，那么这个用户恶意欺诈的可能性就较大。

移动大数据可以帮助P2P公司在一定程度上来验证贷款用户真实工作地点，降低犯罪分子利用高薪工作进行恶意欺诈的风险。

3）欺诈聚集地的识别

恶意欺诈往往具有团伙作案和集中作案的特点。犯罪团伙成员常常会在集中在一个临时地点，雇佣一些人，短时间内进行疯狂作案。

大多是情况下，多个贷款用户在同一个小区居住的概率较低，同时贷款的概率更低。如果P2P平台发现短短几天内，在同一个GPS经纬度，出现了大量贷款请求。并且用户信息很相似，申请者居住在偏远郊区，这些贷款请求的恶意欺诈可能性就较大。P2P公司可以将这些异常行为定义为高风险事件，利用其他的信息进一步识别和验证，降低恶意欺诈的风险。

移动设备的位置信息可以帮助 P2P 公司，识别出出现在同一个经纬度的群体性恶意欺诈事件，降低不良贷款发生概率。

4. 高风险贷款用户的识别

高风险客户也是 P2P 企业的一个风险。高风险客户定义比较广泛，除了信用风险，贷款人的身体健康情况也是一个重要参考。移动大数据的位置信息、安装的 APP 类型、APP 使用习惯，在一定程度上反映了贷款用户的高风险行为。

P2P 企业可以利用移动设备的位置信息，了解过去 3 个月用户的行为轨迹。如果某个用户经常在半夜 2 点出现在酒吧等危险区域，并且经常有飙车行为，这个客户定义成高风险客户的概率就较高。移动 APP 的使用习惯和某些高风险 APP 也可以帮助 P2P 企业识别出用户的高风险行为。

当用户具有以上的危险行为时，其身体健康就面临着较大的威胁，P2P 企业可以参考移动数据，提高将客户列为高风险客户的概率，拒绝贷款或者提前收回贷款。降低用户危险行为导致坏账的风险。

移动大数据在预防互联网恶意欺诈和高风险客户识别方面，已经有了成熟的应用场景。前海征信、宜信、聚信立、闪银已经开始利用 TalkingData 的数据，预防互联网恶意欺诈和识别高风险客户，并取得了较好的效果。移动大数据应用场景正在被逐步挖掘出来，未来移动大数商业应用将更加广阔。

9.5　大数据应用案例之：大吃一惊！大数据下的中国原来是这样的

电影《美国队长 2》中有句台词：21 世纪就是本数据书。如今，大数据越来越被广泛应用。必应搜索通过集成以往的飞机票价刻画出未来票价的走势；Google 利用用户搜索记录判断出美国流感疫情的现状，比疾控中心快一两周；对冲基金通过剖析社交网络推特的数据信息来预测股市的表现……

整合了一些数据分析下的国人衣食住行的真实情况，大数据下的中国或许会令你大吃一惊！

例如，2013 年中国产生的数据总量超过 0.8ZB，相当于 1200 万个中国国家图书馆藏书量；2013 年世界上所储存的数据如果印刷成书，则可以覆盖整个美国 52 次。

1. 超过 2.8 亿的中国人缺乏安全用水

环境保护部今年发布了首个全国性的大规模研究结果。结果显示，我国有 2.5 亿居民的住宅区靠近重点排污企业和交通干道，2.8 亿居民在使用不安全饮用水，如图 9.7 所示。

由于规划和产业布局原因，我国有 1.1 亿居民住宅周边 1 公里范围内有石化、炼焦、火力发电等重点关注的排污企业，1.4 亿居民住宅周边 50m 范围内有交通干道。在大气污染物浓度相同的情形下，我国城市居民暴露于大气污染健康风险是农村居民的 70%。



图 9.7　超过 2.8 亿的中国人缺乏安全用水

2. 中国每年生产 800 亿双筷子

中国传统的筷子本来是我们的骄傲，但是一次性筷子的滥用却成了国人的耻辱。据英国《卫报》2006 年报道，中国每年消耗 450 亿双筷子，如图 9.8 所示。



图 9.8　中国每年生产 800 亿双筷子

情况似乎在今年内变得更糟。最新调查结果显示，我国 2013 年生产了 800 亿双一次性筷子。这需要砍伐 2000 万棵生长了 20 年的大树！如果每双筷子按长度 20 厘米、宽度

1 厘米、厚度 0.5 厘米计算，800 亿双筷子可铺满 363 个天安门广场。

3. 国人平均睡眠时间 7.05 小时南方人更爱熬夜

睡眠是生命中最珍贵的事，我们通过大数据分析发现国人的几个怪现象：南方人比北方人更爱熬夜，单身比恋爱中的人睡得多……

调查报告显示，“中国睡眠指数”的总得分为 66.5 分，较去年的 64.3 分提升了 2.2 分，表明我国居民整体睡眠状况呈现向好发展趋势。但其中超过三成（36.2%）居民的得分低于及格线（60 分），这也说明了国人的睡眠状况两极化趋势渐显：整体来看，人们开始享受舒适的睡眠，但同时也有更多的人饱受睡眠障碍的困扰，如图 9.9 所示。



图 9.9　国人平均睡眠时间 7.05 小时

QQ 大数据发布《网民睡眠质量报告》显示，我国网民平均睡眠时间为 7.05 小时，而一线城市网民睡眠时间最少，仅为 6.95 小时，熬夜用户占到 20.9%。据统计，在即使是在最爱睡的城市呼和浩特，人们的平均睡眠也只有 7.33 小时，没有达到 8 小时。

相比女性，男性熬夜时间更长，较女性高了 4.7 个百分点。而就年龄段而言，90 后的熬夜能力是最长的，达到了人数的 31.5%，果然是年轻气盛啊，如图 9.10 所示。

热恋中的用户平均比单身用户多睡 18 分钟，可见人是个不甘寂寞的物种……

大部分网民半夜不睡觉，都在做什么呢？据统计，44.8%的人在深夜追剧，另外 43.8%的人在深夜中打游戏，大约有 1/4 的人会熬夜看书，煲电话粥则占了 13.8%。看美剧和玩游戏都是时间的一大杀器。

在睡姿统计中，不同地域的人往往会选择不同的睡姿。可能睡姿也一定程度上暴露

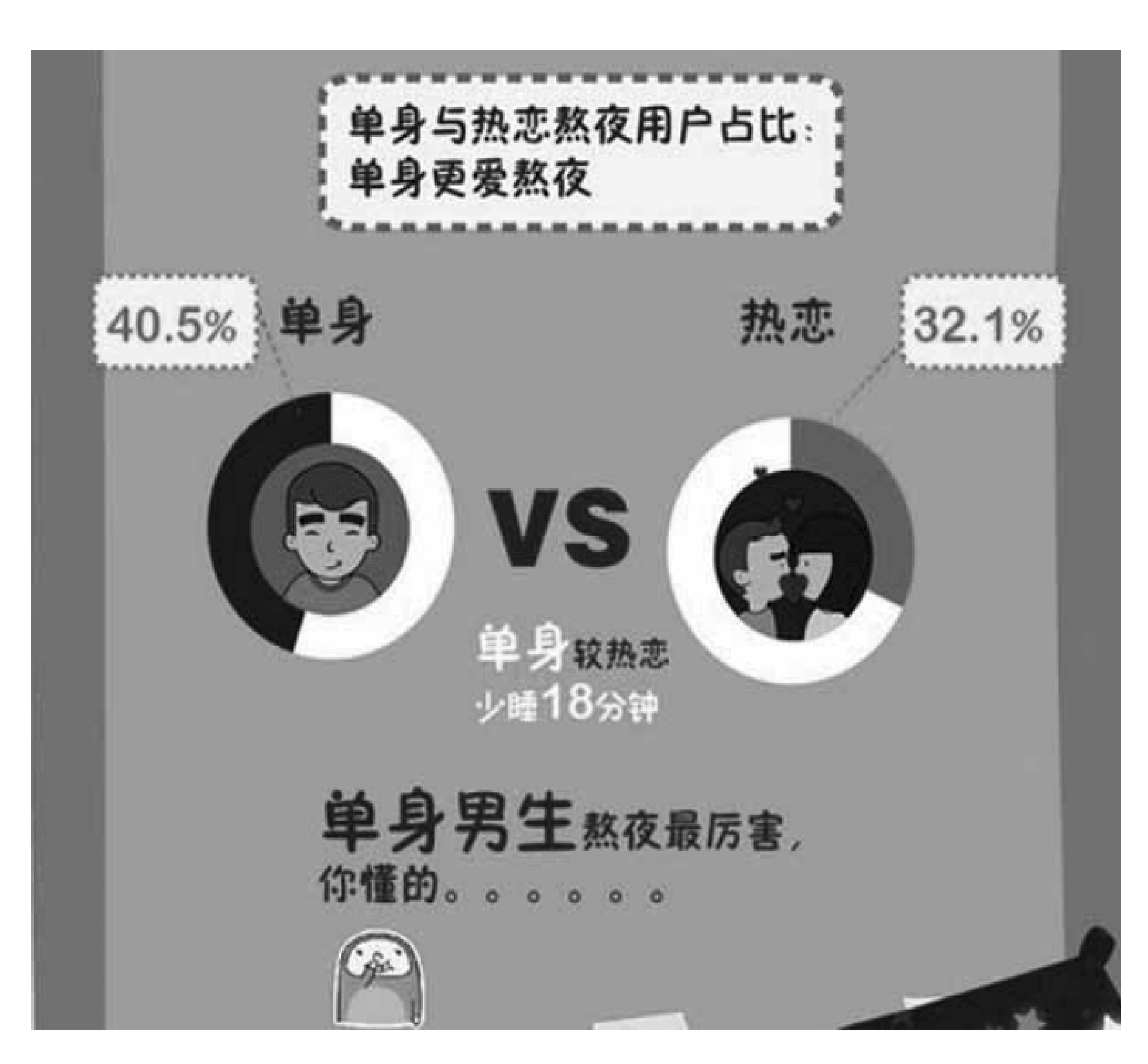


图 9.10 男性熬夜时间更长

了性格，豪放的“东北爷们”最爱熊抱睡，如图 9.11 所示。

由于历史等各方面原因，南北方人在身高饮食等各个方面都有着不小的差异，没想到连睡觉时间都不一样。南方人的熬夜指数比北方大约高 5 个百分点。

从 2010 年至今，手机等电子产品的普及和发展进一步“偷”走了用户的时间。很多人熬夜上网，玩游戏、看小说，对身体有着极大的危害。5 年来，熬夜人数上升了 8%，平均睡眠时间也下降了一小时，由原来的 8.1 小时降到 7.05 小时。

睡眠和健康是直接相关的，在科技爆炸的时代，睡眠时间和质量越来越受人关注，建议国人还是要减少熬夜，保持身体健康。

4. 卫星数据展现雾霾笼罩下的中国

中国的雾霾备受世界瞩目，美国宇航局曾公布 VIIRS(NASA 的 NPP 卫星搭载的可见光红外成像辐射套件)设备拍摄亚洲上空的雾霾画面。清晰地看到，中国华北一带的上空是厚厚的灰色雾霾层，如图 9.12 所示。

通过卫星数据看到雾霾后，我们得利用大数据解决雾霾的问题。有研究机构称，可以根据现有监测站所提供的空气质量数据以及城市里的其他多种数据来源(包括气象情况、交通流量、人员流动趋向、路网结构、人口集中点等)，运用数据挖掘和机器学习技术，对大数据加以充分利用，并在监测信息和对应结果之间建立一个隐式映射，从而可以实时推断出包含细颗粒物信息的城市空气质量数据。

据悉，中国准备在京津冀、长三角和珠三角地区建立雾霾应急减控对策系统，这个系统依托于“天河一号”计算机，可以将采集到的海量空气质量数据进行分析，以对雾霾做出全面的分析及准确的预报，如图 9.13 所示。

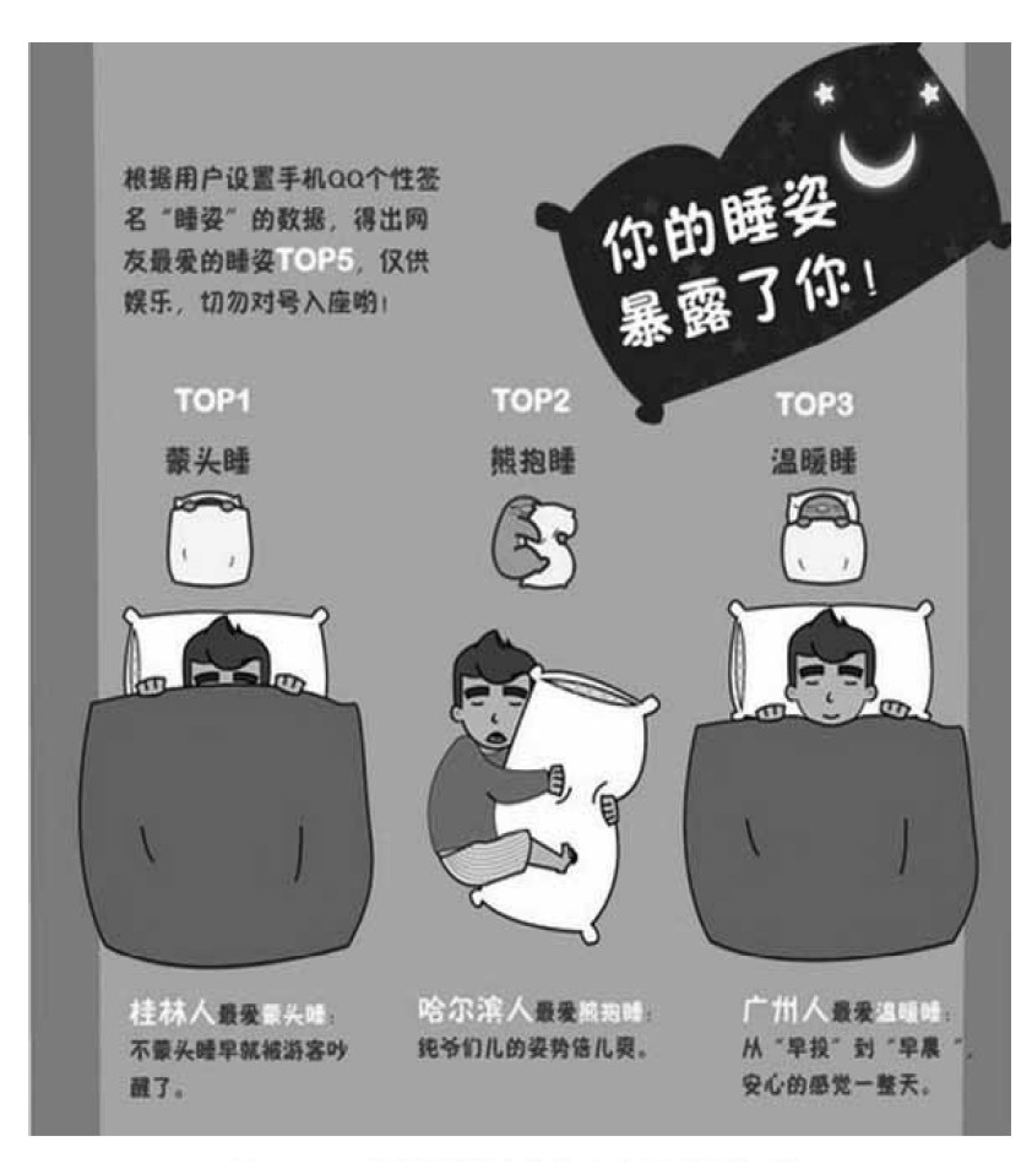


图 9.11　不同地域的人往往会选择不同的睡姿

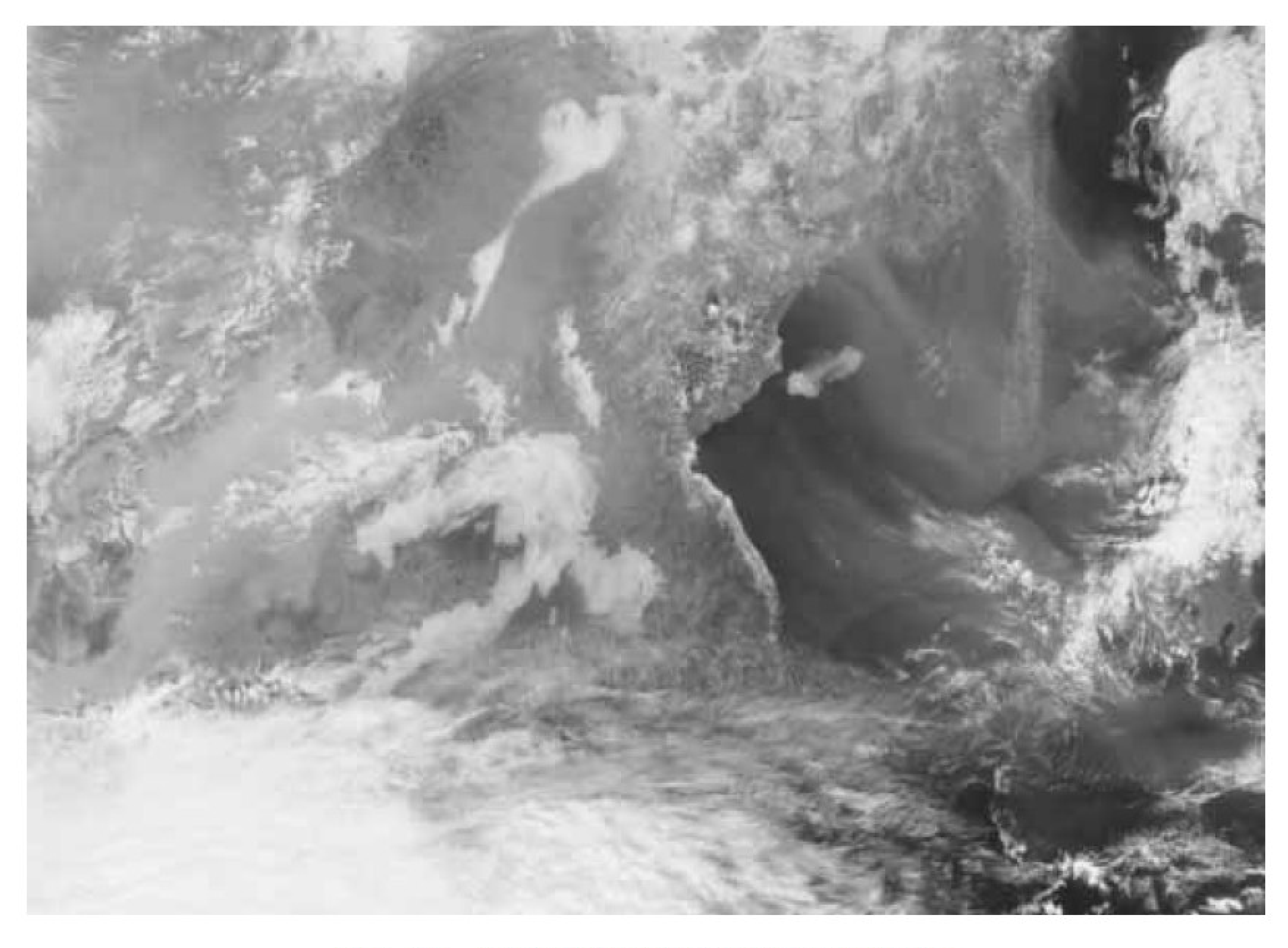


图 9.12　卫星数据展现雾霾笼罩下的中国

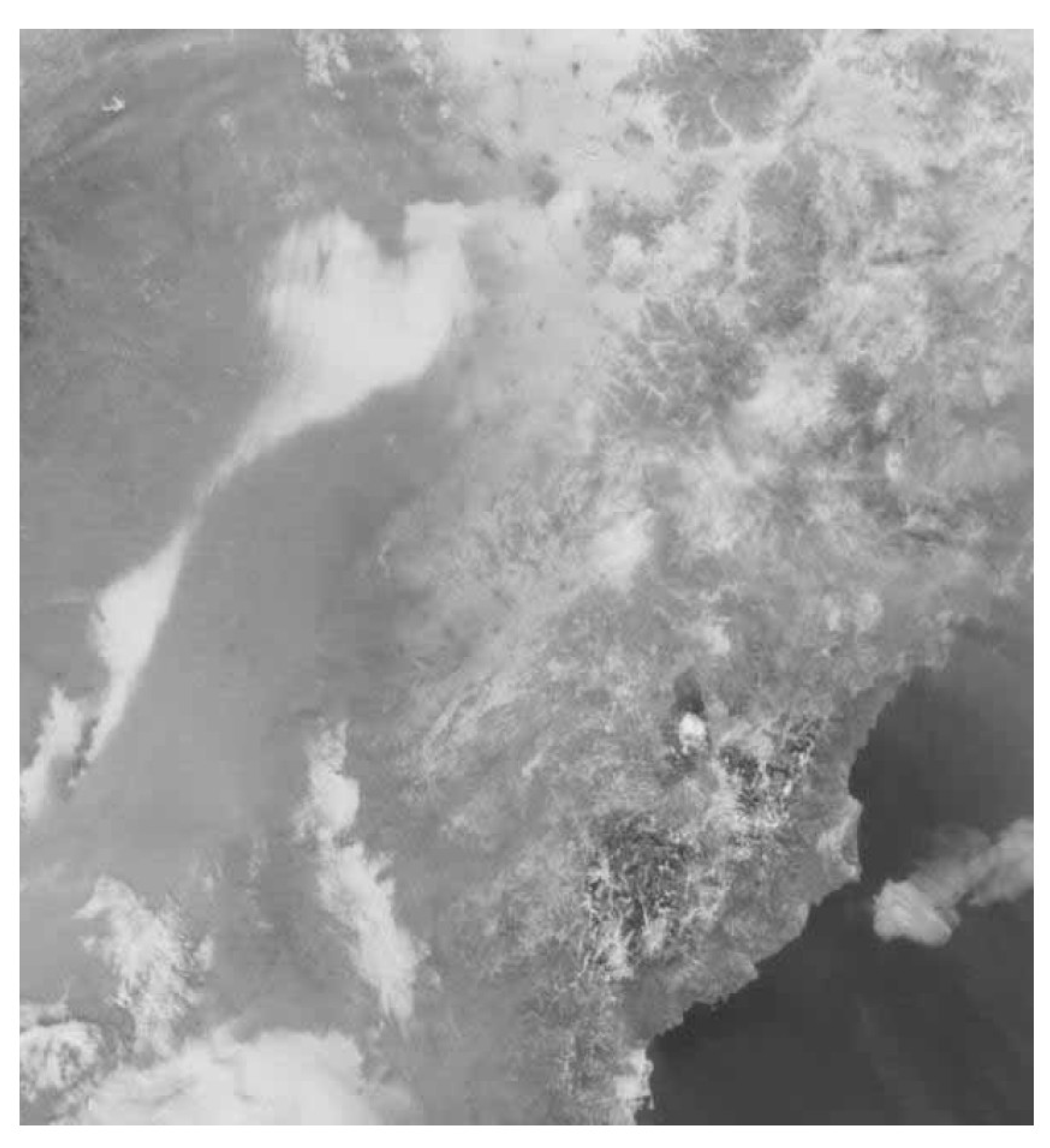


图 9.13　中国局部卫星图展现的雾霾



大数据技术现状及发展展望

第 10 章　大数据技术发展前景

10.1　大数据引发新一代信息技术变革浪潮

大数据领域已经涌现出了大量新的技术，它们成为大数据采集、存储、处理和呈现的有力武器。这些技术下一步将如何发展？它们之中哪些技术将广为流行？又会诞生哪些新的技术？

1. 技术趋向多样化，企业应选择未来会快速普及的技术

目前，大数据相关的技术和工具非常多，给企业提供了更多的选择。在未来，还会继续出现新的技术和工具，如 Hadoop 分发、下一代数据仓库等，这也是大数据领域的创新热点。

那么企业到底该选用什么技术呢？

TDWI(数据仓库研究所)对现有的大部分技术和工具进行了调查，以现在及未来三年内企业接受度和增长率两个维度进行划分，这些技术和工具可分成四类：第 1 类：先进分析法。第 2 类：先进数据可视化。第 3 类：实时化仪表盘。第 4 类：内存数据库。

企业最需要关注的是第 1 类中的技术和工具，它们最有可能成为最佳的实施工具，也代表了大数据技术的发展方向。

2. 基于云的数据分析平台将更趋完善

企业越来越希望能将自己的各类应用程序及基础设施转移到云平台上。就像其他 IT 系统那样，大数据的分析工具和数据库也将走向云计算。

云计算能为大数据带来哪些变化呢？

首先云计算为大数据提供了可以弹性扩展、相对便宜的存储空间和计算资源，使得中小企业也可以像亚马逊一样通过云计算来完成大数据分析，如图 10.1 所示。

其次，云计算 IT 资源庞大、分布较为广泛，是异构系统较多的企业及时准确处理数据的有力方式，甚至是唯一的方式。

当然，大数据要走向云计算，还有赖于数据通信带宽的提高和云资源池的建设，需要确保原始数据能迁移到云环境以及资源池可以随需弹性扩展。

数据分析集逐步扩大，企业级数据仓库将成为主流，未来还将逐步纳入行业数据、政府公开数据等多来源数据，如图 10.2 所示。

当人们从大数据分析中尝到甜头以后，数据分析集就会逐步扩大。目前大部分的企业所分析的数据量一般以 TB 为单位。按照目前数据的发展速度，很快将会进入 PB 时

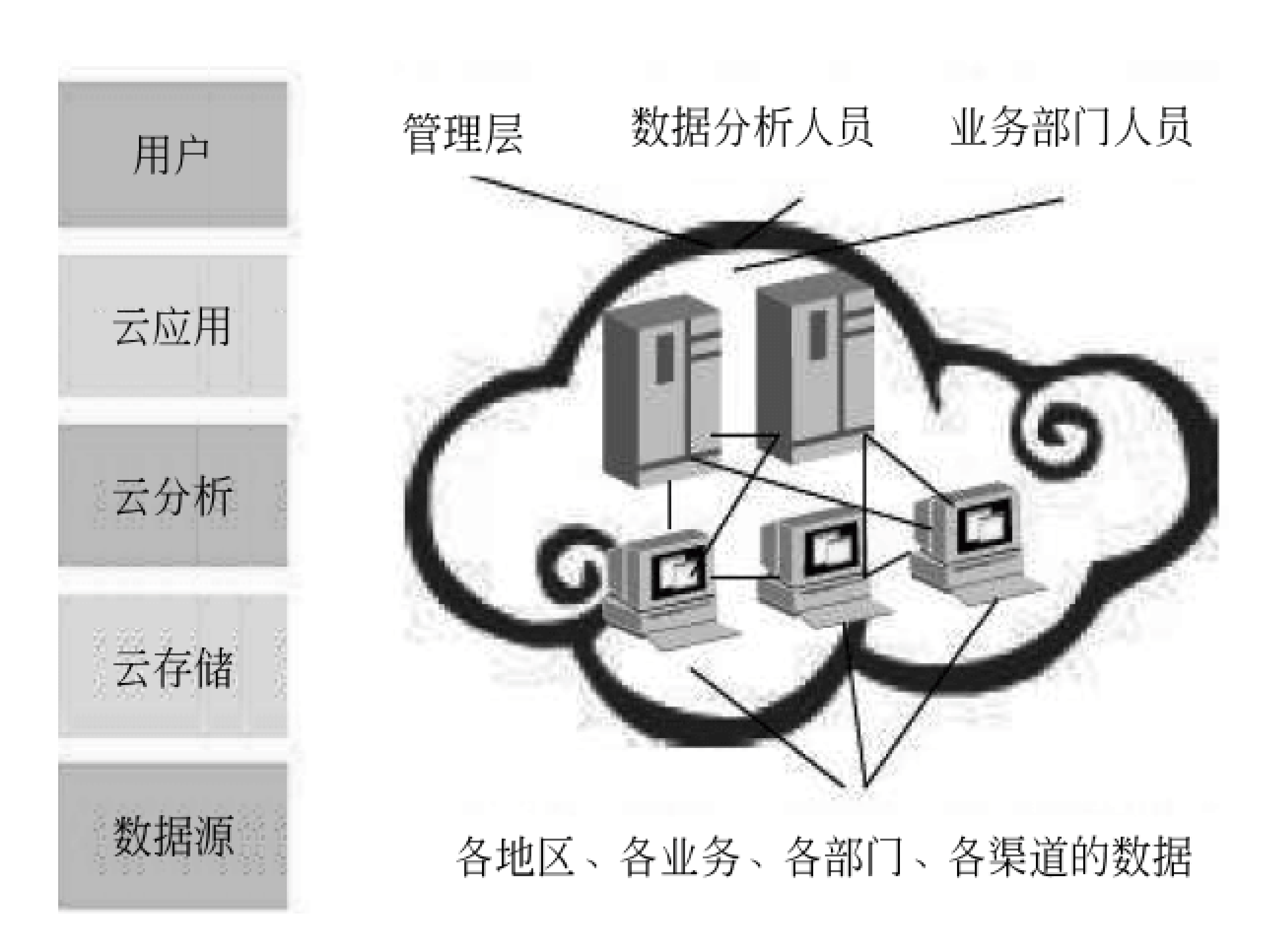


图 10.1 基于云的数据分析平台框架

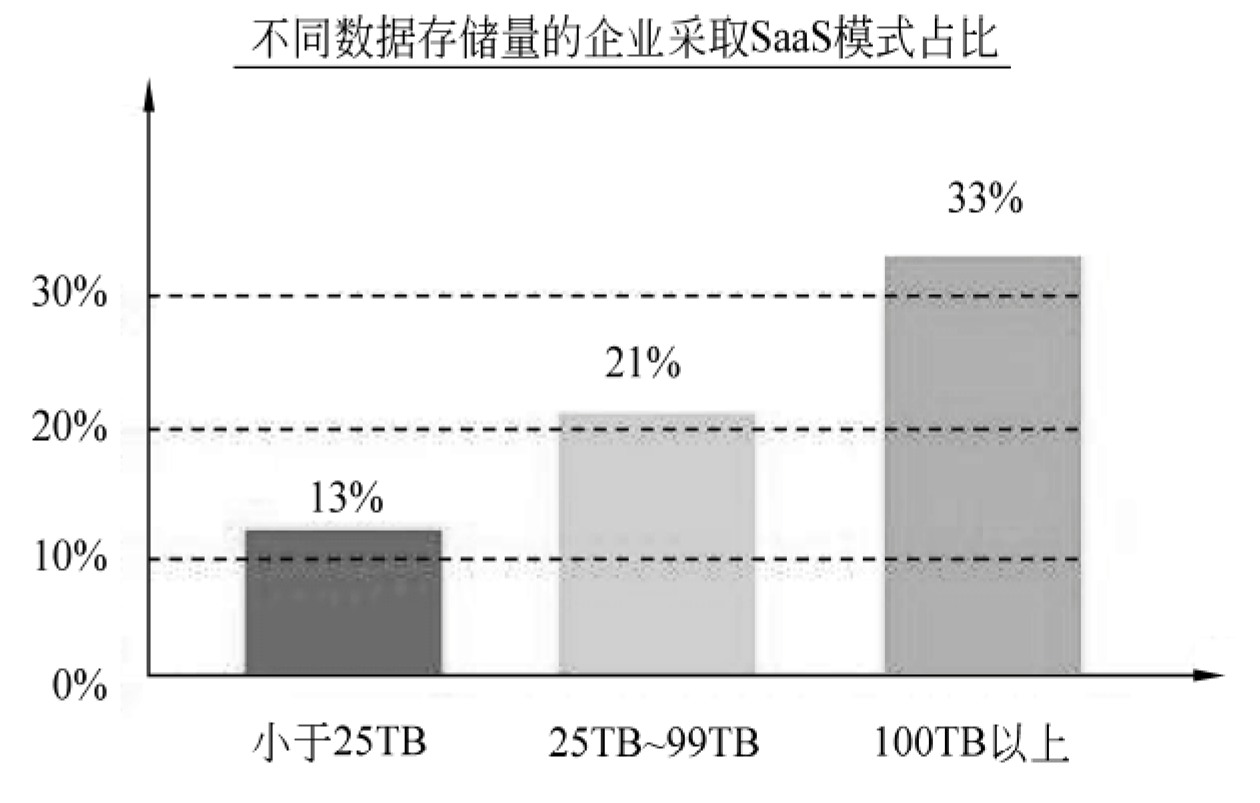


图 10.2 不同数据存储量的企业采取不同的存储模式

代。特别是目前在 100～500TB 和 500＋TB 范围的分析数据集的数量会呈 3 倍或 4 倍增长。

随着数据分析集的扩大,以前部门层级的数据集市将不能满足大数据分析的需求,它们将成为企业级数据库(EDW)的一个子集。根据 TDWI 的调查,如今大概有 2/3 的用户已经在使用企业级数据仓库,未来这一占比将会更高。传统分析数据库可以正常持续,但是会有一些变化,一方面,数据集市和操作性数据存储(ODS)的数量会减少;另一方面,传统的数据库厂商会提升它们产品的数据容量、细目数据和数据类型,以满足大数据分析的需要。

因此,企业内的数据分析将从部门级过渡到企业级,从面向部门需求转向面向企业需求,从而也必将获得比部门视角更大的益处。

需要指出的是,随着政府和行业数据的开放,更多的外部数据将进入企业级数据仓库,使得数据仓库规模更大,数据的价值也越大。

10.2　大数据采集与预处理技术发展前景

根据大数据处理的生命周期，大数据的技术体系通常可以分为大数据采集与预处理、大数据存储与管理、大数据计算模式与系统、大数据分析与挖掘、大数据可视化计算以及大数据隐私与安全等几个方面。

1. 问题与挑战

通常大数据描述了一个对象(物理的或逻辑的)或一个过程的全景式的和全周期的状态，因此，其来源必然是多源的，其形式是多模态的。数据的多源和多模态的不确定性和多样性，必然导致数据的质量存在差异，严重影响到数据的可用性。

由于数据量的大规模性，即使错误数据的相对比例不大，而绝对的错误数据量也是非常可观的。据国际咨询机构调查，全球财富 1000 强企业中 25%以上的企业信息信息系统存在不正确的数据，美国企业信息系统中 1%～30%的数据存在各种错误，美国工业企业由于数据错误而引起的生产事故和决策错误，每年造成 6000 多亿美元的损失。

数据的可用性取决于数据质量。数据质量的定义有很多说法。按照一般的的定义，数据质量包含五种特性：精确性、一致性、完整性、同一性和实效性。

精确性指数据符合规定的精度，不超出误差范围；一致性指数据之间不能存在相互矛盾；完整性指数据的值不能为空；同一性指实体的标识是唯一的；时效性指数据的值反映了实际的状态。此外，考虑到人为因素，还可以要求第六个性质，即真实性，即数据不能是人工伪造的。

2. 主要进展

针对管理信息系统中异构数据库集成技术、Web 信息系统中的实体识别技术和 DeepWeb 集成技术、传感器网络数据融合技术已经有很多研究工作，取得了较大的进展，已经推出了多种数据清洗和质量控制工具，例如，美国 SAS 公司的 Data Flux、美国 IBM 公司的 Data Stage、美国 Informatica 公司的 Informatica Power Center。

但是，针对各种类型、各种应用的大数据的特点，如何保证一致性、精确性、完整性、统一性、时效性、真实性六个性质，并且保证可行的处理效率，还缺乏全面系统的研究，许多新问题有待于发现和解决。

3. 发展趋势

为了保证大数据的可用性，首先必须在数据的源头上把好质量关，做好从原始数据到高质量信息的预处理。具体的关键技术有如下几种。

1）数据源的选择和高质量原始数据的采集方法

用于从可靠的高质量数据源里，获得高质量的原始数据。为了确保数据源的质量，需要建立数据源的质量评估理论模型，包括数据源的综合质量评估和高质量数据源的选择方法。然后，针对各种模态数据的特点，建立高质量多模态数据的获取方法，包括有效的数据采集方法、多模态数据融合算法、数据的保质转换算法、数据精确性和一致性方面的错误校验和纠错、数据完整性方面的缺失值估计、数据的时效性检测、数的真实性验证等。

2）多源数据的实体识别和解析方法

用于识别和合并相同的实体，区分不同的实体。为了高质量的数据集成奠定基础，必须保证数据的实体同一性，解决来自多个数据源的多模态数据的实体识别问题。需要建立多源数据的实体关联模型和识别模型、多源多模态数据的实体自动识别方法、实体识别效果的评估模型等。

3）数据清洗和自动修复方法

根据正确性条件和数据约束规则，清除不合理和错误的数据，对重要的信息进行修复，保证数据的完整性。需要建立数据正确性语义模型、关联模型和数据约束规则、数据错误模型和错误识别学习框架、针对不同错误类型的自动检测和修复算法、错误检测与修复结果的评估模型和评估方法等。

4）高质量的数据整合方法

在数据采集和实体识别的基础上，进而实现数据到信息的高质量整合。需要建立多源多模态信息集成模型、异构数据智能转换模型、异构数据集成的智能模式抽取和模式匹配算法、自动的容错映射和转换模型及算法、整合信息的正确性验证方法、整合信息的可用性评估方法等。

5）数据演化的溯源管理

用于对数据的演化过程进行跟踪和记录，以保证和控制数据的质量。需要建立世系模型及其追踪技术，主要包括时空、多粒度、多路径和不确定的海量信息演化的演化模型和演化描述方法、演化模式的正向性评估模型与方法、演化的可逆性判定与近似求解算法、分布式、多粒度、概率化的世系追踪技术等。

总之，大数据的采集和预处理是大数据的源头，在源头上把好质量关，对大数据的后续处理和分析至关重要。因此，对大数据的使用者、研究者、开发者以及上级主管部门，提出如下建议：

(1) 提高用户对大数据可用性的重要性的认识，切实开展大数据质量控制，确保大数据处理和分析结果的正确性。

(2) 针对大数据质量控制面临的挑战性问题，学术界应加强对大数据可用性评估和保证的关键技术的研究和开发。

(3) 大数据的质量控制具有广泛的需求和巨大的市场前景，工业界应注重大数据可用性的评估，加强数据质量保证软件的开发和推广。

(4) 建议政府有关部门尽快建立关于大数据可用性(数据质量)的标准，保证大数据的统一质量，有效保证大数据的利用价值。

10.3 大数据存储与管理技术发展前景

1. 问题与挑战

大数据给存储系统带来了三个方面的挑战：

(1) 存储规模大，通常达到 PB(1000TB)甚至 EB(1000PB)量级。

(2) 存储管理复杂，需要兼顾结构化、非结构化和半结构化的数据。

（3）数据服务的种类和水平要求高，换言之，上层应用对存储系统的性能、可靠性等指标有不同的要求，而数据的大规模和高复杂度放大了达到这些指标的技术难度。这些挑战在存储领域并不是新问题，但在大数据背景下，解决这些问题的技术难度成倍提高，数据的量变终将引起存储技术的质变。

大数据环境下的存储与管理软件栈，需要对上层应用提供高效的数据访问接口，存取 PB 甚至 EB 量级的数据，并且能够在可接受的响应时间内完成数据的存取，同时保证数据的正确性和可用性；对底层设备，存储软件栈需要充分高效的管理存储资源，合理地利用设备的物理特性，以满足上层应用对存储性能和可靠性的要求。在大数据带来的新挑战下，要完成以上这些要求，需要更进一步的研究存储与管理软件技术。

2. 主要进展

根据为上层应用提供的访问接口和功能侧重不同，存储与管理软件主要包括文件系统和数据库；在大数据环境下，目前最适用的技术是分布式文件系统、分布式数据库以及访问接口和查询语言。

1）分布式文件系统

分布式文件系统所管理的数据存储在分散的设备或结点上，存储资源通过网络连接。用分布式文件系统对大数据进行存储与管理，目前的研究主要涉及以下几个关键的技术：

（1）高效元数据管理技术。

大数据应用下，元数据的规模也非常大，元数据的存取性能是整个分布式文件系统性能的关键。常见的元数据管理可以分为集中式和分布式元数据管理架构。集中式元数据管理架构采用单一的元数据服务器，优点是实现简单，但存在单点故障等问题。分布式元数据管理架构则将元数据分散在多个结点上，从而解决了元数据服务器性能瓶颈问题，提高了可扩展性，但实现复杂，并引入了元数据一致性的问题。

此外，还有一种无元数据服务器的分布式架构，使用在线算法组织数据，不需要专用的元数据服务器。但是该架构对数据一致性的保证很困难，实现复杂。文件目录遍历操作的效率低下，并且缺乏文件系统全局监控管理功能。

（2）系统弹性扩展技术。

大数据环境下，数据规模和复杂度的增加往往非常迅速，所以按需扩展系统规模是十分必要的。实现存储系统的高可扩展性首先要解决两个方面的重要问题：元数据的分配和数据的透明迁移。前者主要通过静态子树划分和动态子树划分技术实现，后者则侧重数据迁移算法的优化。

此外，大数据存储系统规模庞大，结点失效率高，因此还需要实现一定程度上的自适应管理功能。系统必须能够根据数据量和计算的工作量估算所需要的结点个数，并动态地将数据在结点间迁移，以实现负载均衡；同时，结点失效时，数据必须可以通过副本等机制进行恢复，不能对上层应用产生影响。

（3）存储层级内的优化技术。

构建存储系统时，需要基于成本和性能来考虑，因此存储系统通常采用多层不同性价比的存储器件组成存储层次结构。大数据的规模大，因此构建高效合理的存储层次结构，

可以在保证系统性能的前提下，降低系统能耗和构建成本。利用数据访问局部性原理，可以从两个方面对存储层次结构进行优化。

从提高性能的角度，通过分析应用特征，识别热点数据并对其进行缓存或与预取，通过高效的缓存预取算法和合理的缓存容量配比，以提高访问性能。从降低成本的角度，采用信息生命周期管理方法，将访问频率低的冷数据迁移到低速廉价存储设备上，可以在小幅牺牲系统整体性能的基础上，大幅降低系统的构建成本和能耗。

(4) 针对应用和负载的存储优化技术。

传统数据存储模型需要支持尽可能多的应用，因此需要具备较好的通用性。大数据具有大规模、高动态及快速处理等特性，通用的数据存储模型通常并不是最能提高应用性能的模型，而大数据存储系统对上层应用性能的关注远超过对通用性的追求。针对应用和负载来优化存储，就是将数据存储与应用耦合，放宽 POSIX 接口，简化或扩展分布式文件系统的功能，根据特定应用、特定负载、特定的计算模型对文件系统进行定制和深度优化，使应用达到最佳性能。这类优化技术在 Google、Facebook 等互联网公司的内部存储系统上，管理超过 PB 级的大数据，能够达到非常高的性能。

(5) 针对存储器件特性的优化技术。

随着新型存储器件的发展和成熟，Flash、PCM 等逐渐开始在存储层级中占据一席之地，存储软件栈也随之开始逐渐发生变化。以 Flash 为例，起初各厂商通过闪存转换层 FTL 对新型存储器进行封装，以屏蔽存储器件的特性，适应存储软件栈的现有接口。但是随着 Flash 的普及，产生了许多针对应用对 FTL 进行的优化，以及针对 Flash 特性进行定制的文件系统，甚至有去掉 FTL 这层冗余直接操作 Flash 的存储解决方案。

传统的本地文件系统，包括分布式文件系统，是否能够与新型存储器件耦合，最大程度地利用这些存储器件新特性上的优势，需要存储软件开发者重新审视存储软件栈，去除存储软件栈的冗余，甚至需要修复一些不再合适的部分。

2) 分布式数据库

大数据时代企业对数据的管理、查询及分析的需求变化促生了一些新的技术的出现。需求的变化主要集中在数据规模的增长、吞吐量的上升、数据类型以及应用多样性的变化上。数据规模和吞吐量的增长需求对传统的关系型数据库管理系统在并行处理，事务特性的保证，互联协议的实现，资源管理以及容错等各个方面带来了很多挑战。而数据类型以及应用的多样性带来了为了支持不同应用的数据管理系统。

(1) 事务性数据库。

这类数据库主要包括 NoSQL 和 NewSQL。NoSQL(“Not Only SQL”或者“Not Relational”)系统往往通过放松对事务 ACID 语义的方法来增加系统的性能以及可扩展性(CAP 定理)。NoSQL32 系统往往具有以下几个特征：

① 非关系数据模型，比如键值存储等。

② 对简单操作比如键值查询的水平可扩展性，往往不支持 SQL 全集。

③ 在多个结点中分割和复制数据的能力。

④ 弱并发一致性语义(比如最终一致性)。

⑤ 充分利用分布式索引和内存。

根据管理数据的模式分类，NoSQL 系统可以分为三类：键值系统、文档存储系统以及图数据库。键值系统的代表性系统包括 BigTable、Dynamo、HBase、Gemfire、Redis、Cassandra，文档存储系统的代表包括 MongoDB 和 Couchbase，图数据库的代表是 Neo4j 等等。

NoSQL 系统通过对事务语义的放松达到系统的可扩展性，但是把一致性的维护交由用户来管理，这对很多对一致性要求不高的应用来说是足够的。但是如果应用需要保证一致性，对开发人员来说就很困难了。NewSQL 就是在这样的背景下诞生的。NewSQL 系统可以在提供类似 NoSQL 的可扩展性的同时保证事务 ACID 属性，并且提供 SQL 用户接口。NewSQL 系统通常可以分为两类。

① 通用数据库：这类系统保持传统分布式数据库的功能，但是在设计分布式体系架构时充分考虑了大规模高吞吐系统的特性。这类系统的典型代表是 Spanner 和 NuoDB。

② 基于内存的数据库：这类系统基本上针对的是高吞吐短小事务，不再采用传统的关系型数据库设计。这类数据库的典型代表是 SQLFire 和 VoltDB。

（2）分析型数据库。

分析型数据库在大数据时代也呈现了一种百家争鸣的局面。自从 MapReduce 被提出以及 Hadoop33 的流行，出现了多家针对 Hadoop 的 SQL 分析引擎，代表性系统包括 Hive、HAWQ、Impala 和 Hadapt。

Hive34 是一个基于 MapReduce 的 SQL 引擎。基本原理是接受 SQL，解析 SQL，然后把 SQL 语句翻译成多个 MapReduce 的任务，通过 MapReduce 来实现基本的 SQL 操作。因为 Hive 基于 MapReduce，所以它把容错、执行以及资源管理的工作都交给了 MapReduce 框架，其特点是简单与易于实现。但是它也有一些不可避免的缺陷，包括对标准 SQL 以及实时查询的支持，难于优化带来的查询性能低下，并且很难充分利用整个集群的资源，从而导致并发吞吐量较低。

HAWQ35（Hadoop with Query）是 Hadoop 领域与 SQL 兼容的大规模数据分析引擎。HAWQ 继承了 Hadoop 与 MPP 大规模数据库分析引擎的优点，实现了 HDFS 分布式存储与 MPP 执行引擎的结合。HAWQ 实现了 MPP 基于统计的优化器，支持数百万连接的网络互联协议、数据的多级划分与存储和高效的执行引擎。

其特点是与各种 BI 工具的兼容，实时查询的支持，以及与基于 MapReduce 系统的性能优势。

Impala 和 Hadapt 是另外两个基于 Hadoop 的 SQL 引擎。其基本的出发点也是把 MPP 的技术引入 Hadoop。但目前还不是很成熟。

3）访问接口和查询语言

大数据系统的访问接口和查询语言取决于系统的存储模型。传统的 MPP 数据库都使用关系模型，其查询语言为标准的 SQL。而图数据库有自己的查询语言，可以实现子图匹配、路径查询等功能。

Hadoop 本身使用的是 HDFS，MapReduce 编程接口可以作为其访问接口。构建在 Hadoop 之上的类数据库系统则提供各自存储模型所对应的查询语言和访问接口。例如，HBase 提供 API，用于对数据表进行 key-value 形式的查询和增删改操作。

Hive 则提供称为 HiveQL 的查询语言,用于对关系表进行查询,HiveQL 同 SQL 非常相似,并附带一些 SQL 未提供的功能。为了方便对 hadoop 的使用,一系列的查询语言和附加访问接口被提出。

Pig 是一种基于 MapReduce 的编程平台,它的访问语言 Pig Latin 是介于 SQL 和过程式程序设计语言之间的语言,结合了 SQL 申明式(declarative)语言的优势以及过程式程序设计的灵活性,得到了众多程序设计者的青睐。

Sqoop 是一种用于在关系数据库和 hadoop 之间进行数据迁移的命令式语言。Mahout 则是构建在 hadoop 之上的机器学习引擎,也拥有自己的一套访问接口。

3. 发展趋势

大数据给存储系统的发展趋势是实时/流式大数据存储与处理。

随着业务的增长,业界对大数据的速度(Velocity)维度越来越关注,过去需要几天或者几个小时才能回答的问题现在期望在几分钟、几秒甚至毫秒内得到解决。实时流数据存储和处理技术将会越来越多地被研究和开发。实时流式大数据的处理在很多方面和分布式系统在原理上有很多相似之处,然而也有其独特需求。

实时流数据处理系统包括流数据的实时存储和流数据的实时计算。流数据存储指的是快速高效的存储流式数据到数据库、数据仓库或者数据池中;流数据的实时计算注重对流数据的快速高效处理、计算和分析。

1) 数据流加载

实时流式大数据系统中,数据通常以流的方式进入系统。如何高效且可靠地将数据加载到大数据存储系统中成为流式大数据系统实现低延迟处理的基础。此外能够重新处理数据流中的数据也是一个很有价值的特性。

2) 复杂事件处理(CEP)

数据流中的数据源是多种多样的,数据的格式也是多种多样,而数据的转换、过滤和处理逻辑更是千变万化,因而需要强大而又灵活的复杂事件处理引擎来适应各种场景下的需求。

3) 高可用性

数据通过复杂处理引擎和流计算框架时,通常会经过很多步骤和结点,而其中任何一步都有出错的可能,为了保证数据的可靠性和精准投递,系统需要具有容错和去重能力。

4) 流量控制和缓存

整个流系统可能有若干个模块,每个模块的处理能力和吞吐量差别很大,为了实现总体高效的数据处理,系统需要对流量进行控制和动态结点增加和删除的能力。当数据流入大于流出的速度时,还需要有一定的缓存能力,如果内存不足以缓存快速流入的数据时,需要能够持久化到存储层。

目前市场上已经出现了多种大数据实时处理技术,它们各有不同的侧重点,例如数据传输技术有 Flume、Scribe、Kafka、Sqoop 等,计算框架有 Storm、S4、Spark 等。基于 Hadoop 的 SQL 处理引擎有 Impala、HAWQ 等。另外还有一些产品在大数据流计算框架之上提供分析即服务,例如 Cetas。大数据的实时存储与处理还有很多需要研究和解

决的问题。

10.4　大数据计算模式与系统技术发展前景

1. 问题与挑战

为了能更清晰地理解不同的大数据计算模式，首先需要梳理出大数据处理中主要的数据特征和计算特征维度，在此基础上进一步梳理目前出现的各种重要和典型的大数据计算模式。大数据处理包括以下典型的特征和维度。

1）数据结构特征

根据数据结构特征大数据可分为结构化/半结构化数据处理与非结构化数据处理。

2）数据获取处理方式

按照数据获取方式，大数据可分为批处理与流式计算（streaming）方式。

3）数据处理类型

从数据处理类型来看，大数据处理可分为传统的查询分析计算和复杂的数据挖掘分析计算。

4）实时性或响应性能

从数据计算响应性能角度看，大数据处理可分为实时/准实时与非实时计算，或者是联机（online）计算与线下（offline）计算。流式计算通常属于实时计算，查询分析类计算通常也要求具有高响应性能，而批处理和复杂数据挖掘计算通常属于非实时或线下计算。

5）迭代计算

现实的数据处理中有很多计算问题需要大量的迭代计算（如一些机器学习算法），为此需要提供具有高效的迭代计算能力的计算模式。

6）数据关联性

MapReduce 适用于处理数据关系较为简单的计算任务，但社会网络等具有复杂数据关系的计算任务则需要研究和使用图数据计算模式。

7）并行计算体系结构特征

由于需要支持大规模数据的存储计算，大数据处理通常需要使用基于集群的分布式存储与并行计算体系结构和硬件平台。此外，为了克服传统的 MapReduce 框架在计算性能上的缺陷，人们从体系结构层面上提出了内存计算模式。

2. 主要进展

根据大数据处理多样性的需求和以上不同的特征维度，目前出现了多种典型和重要的大数据计算模式。与这些计算模式相适应，出现了很多对应的大数据计算系统和工具。由于单纯描述计算模式比较抽象和空洞，因此，在描述不同计算模式时，将同时给出相应的典型计算系统和工具，这将有助于对计算模式的理解以及对技术发展现状的把握，并进一步有利于在实际大数据处理应用中对合适的计算技术和系统工具的选择使用。

1）大数据查询分析计算模式与典型系统

由于行业数据规模的增长已大大超过了传统的关系数据库的承载和处理能力，因此，目前需要尽快研究并提供面向大数据存储管理和查询分析的新的技术方法和系统，尤其要解决在数据体量极大时如何能够提供实时或准实时的数据查询分析能力，满足企业日常的经营管理需求的问题。然而，大数据的查询分析处理具有很大的技术挑战，在数量规模较大时，即使采用分布式数据存储管理和并行化计算方法，仍然难以达到关系数据库处理中小规模数据时那样的秒级响应性能。

大数据查询分析计算的典型系统包括 Hadoop36 下的 HBase 和 Hive、Facebook 开发的 Cassandra、Google 公司的 Dremel、Cloudera 公司的实时查询引擎 Impala；此外，为了实现更高性能的数据查询分析，还出现了不少基于内存的分布式数据存储管理和查询系统，如 UC Berkeley AMPLab 的基于内存计算引擎 Spark 的数据仓库 Shark、SAP 公司的 Hana 等，如表 10.1 所示。

表 10.1　典型大数据计算模式与系统

大数据计算模式与系统	代 表 产 品
大数据查询分析计算	HBase，Hive，Cassandra，Impala，Shark，Hana
批处理计算	Hadoop MapReduce，Spark
迭代计算	HaLoop，iMapReduce，Twister，Spark
图计算	Pregel，Giraph，Trinity，PowerGraph，GraphX
流式计算	Scribe，Flume，Storm，S4，Spark Steaming
内存计算	Dremel，Hana，Spark

2）批处理计算模式与典型系统

最适合于完成大数据批处理的计算模式是 MapReduce。MapReduce 是一个单输入、两阶段（Map 和 Reduce）的数据处理过程。首先，MapReduce 对具有简单数据关系、易于划分的大规模数据采用“分而治之”的并行处理思想；然后将大量重复的数据记录处理过程总结成 Map 和 Reduce 两个抽象的操作；最后 MapReduce 提供了一个统一的并行计算框架，把并行计算所涉及的诸多系统层细节都交给计算框架去完成，以此大大简化了程序员进行并行化程序设计的负担。

MapReduce 的简单易用性使其成为目前大数据处理最为成功、最广为接受使用的主流并行计算模式。在开源社区的努力下，开源的 Hadoop 系统目前已发展成为较为成熟的大数据处理平台，并已发展成一个包括众多数据处理工具和环境的完整的生态系统。目前几乎国内外的各个著名 IT 企业都在使用 Hadoop 平台进行企业内大数据的计算处理。Spark 也是一个批处理系统，其性能方面比 Hadoop MapReduce 有很大的提升，但是其易用性方面目前仍不如 Hadoop MapReduce。

3）流式计算模式与典型系统

流式计算是一种高实时性的计算模式，需要对一定时间窗口内应用系统产生的新数据完成实时的计算处理，避免造成数据堆积和丢失。很多行业的大数据应用，如电信、电力、道路监控等行业应用以及互联网行业的访问日志处理，都同时具有高流量的流式数据

和大量积累的历史数据，因而在提供批处理数据模式的同时，系统还需具备高实时性的流式计算能力。流式计算的一个特点是数据运动、运算不动，不同的运算结点常常绑定在不同的服务器上。

Facebook 的 Scribe 和 Apache 的 Flume 都提供了机制来构建日志数据处理流图。而更为通用的流式计算系统是 Twitter 公司的 Storm37、Yahoo 公司的 S4 以及 UCBerkeley AMPLab 的 Spark Steaming。

4）迭代计算模式与典型系统

为了克服 Hadoop MapReduce 难以支持迭代计算的缺陷，业界和学术界对 Hadoop MapReduce 进行了不少改进研究。HaLoop 把迭代控制放到 MapReduce 作业执行的框架内部，并通过循环敏感的调度器保证前次迭代的 Reduce 输出和本次迭代的 Map 输入数据在同一台物理机上，以减少迭代间的数据传输开销。

iMapReduce 在这个基础上保持 Map 和 Reduce 任务的持久性，规避启动和调度开销；而 Twister 在前两者的基础上进一步引入了可缓存的 Map 和 Reduce 对象，利用内存计算和 pub/sub 网络进行跨结点数据传输。

目前，一个具有快速和灵活的迭代计算能力的典型系统是 UC Berkeley AMPLab 的 Spark，其采用了基于分布式内存的弹性数据集模型实现快速的迭代计算。

5）图计算模式与典型系统

社交网络、Web 链接关系图等都包含大量具有复杂关系的图数据，这些图数据规模常达到数十亿的顶点和上万亿的边数。这样大的数据规模和非常复杂的数据关系，给图数据的存储管理和计算分析带来了很大的技术难题。用 MapReduce 计算模式处理这种具有复杂数据关系的图数据通常不能适应，为此，需要引入图计算模式。

大规模图数据处理首先要解决数据的存储管理问题，通常大规模图数据也需要使用分布式存储方式。但是，由于图数据的数据关系很强，分布存储就带来了一个重要的图划分问题(Graph Partitioning)。在有效的图划分策略下，大规模图数据得以分布存储在不同结点上，并在每个结点上对本地子图进行并行化处理。

与任务并行和数据并行的概念类似，由于图数据并行处理的特殊性，人们提出了一个新的“图并行”(Graph Parallel)的概念。目前已经出现了很多分布式图计算系统，其中较为典型的系统包括 Google 公司的 Pregel、Facebook 对 Pregel 的开源实现 Giraph、微软的 Trinity、Berkeley AMPLab 的 GraphX，以及 CMU 的 GraphLab 以及由其衍生出来的目前性能最快的图数据处理系统 PowerGraph。

6）内存计算模式与典型系统

Hadoop MapReduce 为大数据处理提供了一个很好的平台。然而，由于 MapReduce 设计之初是为大数据线下批处理而设计的，随着很多需要高响应性能的大数据查询分析计算问题的出现，MapReduce 其在计算性能上往往难以满足要求。随着内存价格的不断下降以及服务器可配置的内存容量的不断提高，用内存计算完成高速的大数据处理已经成为大数据计算的一个重要发展趋势。Spark 则是分布内存计算的一个典型的系统，SAP 公司的 Hana 则是一个全内存式的分布式数据库系统。

3. 发展趋势

近几年来,随着大数据处理和应用需求急剧增长,同时也由于大数据处理的多样性和复杂性,针对以上的典型的大数据计算模式,学术界和业界不断研究推出新的或改进已有的计算模式和系统工具平台,目前主要有以下三方面的重要发展趋势和方向。

(1) 主流的 Hadoop 平台改进后将与其他计算模式和平台共存。由于 MapReduce 当初的设计目标主要是针对具有简单数据关系的大数据线下批处理,使得它在系统构架和计算性能上存在不少不足之处,难以适用于那些具有复杂数据关系和复杂计算模式(如迭代计算、图计算等)的大数据处理计算任务。但尽管如此,由于 Hadoop 生态系统已发展成为目前最主流的大数据处理平台,并得到广泛的使用。

考虑到兼容性,目前业界和学术界并不会完全抛弃 Hadoop 平台,而是试图不断改进和发展现有的平台,增加其对各种不同大数据处理问题的适用性。Hadoop 社区正努力扩展现有的计算模式框架和平台,以便能解决现有版本在计算性能、计算模式、系统构架和处理能力上的诸多不足,这正是目前 Hadoop 2.0 新版本"YARN"的努力目标。目前不断有新的计算模式和计算系统出现,预计今后相当长一段时间内,Hadoop 平台将与各种新的计算模式和系统共存,并相互融合,形成新一代的大数据处理系统和平台。

(2) 混合计算模式将成为满足多样性大数据处理和应用需求的有效手段。现实世界中大数据应用复杂多样,可能会同时包含不同特征的数据和计算,在这种情况下单一的计算模式多半难以满足整个应用的需求,因此需要考虑不同计算模式的混搭使用。

混合计算模式可体现在两个层面。一是传统并行计算所关注的体系结构与低层并行程序设计语言层面计算模式的混合,例如,在体结构层,可根据大数据应用问题的需要搭建混合式的系统构架,如 MapReduce 集群+GPU-CUDA 的混合,或者 MapReduce 集群+基于 MIC(Intel Xeon Phi 众核协处理系统)的 OpenMP/MPI 的混合模型。

混合模式的另一个层面是大数据处理高层计算模式的混合。比如,一个大数据应用可能同时需要提供流式计算模式以便接受和处理大量流式数据,提供基于 SQL 或 NoSQL 的数据查询分析能力以便进行日常的数据查询分析,提供线下批处理和迭代计算已完成基于机器学习的深度数据挖掘分析。

一些大数据计算任务可能还涉及复杂图计算或者间接转化为图计算问题。因此,很多大数据处理问题将需要混合使用多种计算模式。此外,为了提高计算性能,各种计算模式还可以与内存计算模式混合,实现高实时性的大数据查询和计算分析。

混合计算模式之集大成者当属 UCBerkeley AMPLab 的 Spark 系统,其涵盖了几乎所有典型的大数据计算模式,包括迭代计算、批处理计算、内存计算、流式计算(Spark Streaming)、数据查询分析计算(Shark)以及图计算(GraphX)。

Spark 提供了一个强大的内存计算引擎,实现了优异的计算性能,同时还保持与 Hadoop 平台的兼容性。因此,随着系统的不断稳定和成熟,Spark 有望成为与 Hadoop 共存的新一代大数据处理系统和平台。

(3) 内存计算将成为高实时性大数据处理的重要技术手段和发展方向。Hadoop 在处理大数据时计算性能不高、难以满足实时性或高响应性计算任务的要求,为此,人们一

直努力改进 Hadoop 的计算性能。但是，在现有 Hadoop 平台面向大数据线下处理的基本构架和工作机制下，性能的改进和提升空间非常有限，难以逾越计算性能低下的障碍；而随着大数据的规模不断扩大，这个问题将越来越为突出。

为此，目前已经逐步形成一个基本共识，即随着内存成本的不断降低，内存计算将成为最终跨越大数据计算性能障碍、实现高实时高响应计算的一个最有效技术手段。因此，目前越来越多的研究者和开发者在关注基于内存计算的大数据处理技术，不断推出各种基于内存计算的计算模式和系统。

内存计算是一种在体系结构层面上的解决方法，因此，它可以与各种不同的计算模式相结合，从基本的数据查询分析计算，到批处理和流式计算，再到迭代计算和图计算，都可以基于内存计算加以实现，因此我们可以看到各种大数据计算模式下都有基于内存计算实现的系统，比较典型的系统包括 SAP 的 Hana 内存数据库、微软的图数据计算系统 Trinity、UC Berkeley AMPLab 的 Spark 等。

由于优异的计算性能，内存计算将成为今后高实时性大数据处理的重要技术手段和发展方向。

10.5　大数据分析与挖掘技术发展前景

1. 问题与挑战

在大数据时代，不同领域不同格式的数据从生活的各个领域涌现出来。大数据往往含有噪声，具有动态异构性，是相互关联和不可信的。尽管含有噪声，大数据往往比小样本数据更有价值。这是因为从频繁模式和相关性分析得到的一般统计量通常会克服个体的波动，会发现更多可靠的隐藏的模式和知识。另一方面，互相连接的大数据形成大型异构信息网。通过信息网，冗余的信息可用于弥补数据缺失所带来的损失，可用于交叉核对数据的不一致性，进一步验证数据间的可信关系，并发现数据中隐藏的关系和模型。

数据挖掘需要集成的、经过清洗的、可信的、可高效访问的数据，需要描述性查询和挖掘界面，需要可扩展的挖掘算法以及大数据计算环境。与此同时，数据挖掘本身也可以用来提高数据质量和可信度，帮助理解数据的语义，提供智能的查询功能。只有能够鲁棒地进行大数据分析，大数据的价值才能发挥出来。另一方面，从大数据得出的知识有助于纠正错误，并消除歧义。

大数据环境下的分析和挖掘方法与传统的小样本统计分析有着根本的不同，并面临如下挑战：

1）数据量的膨胀

随着数据生成的自动化以及数据生成速度的加快，数据分析需要处理的数据量急剧膨胀。一种处理大数据的方法是使用采样技术，通过采样，可以把数据规模变小，以便利用现有的技术手段进行数据管理和分析。

然而在某些应用领域，采样将导致信息的丢失，比如 DNA 分析等。在明细数据上进行分析，意味着需要分析的数据量将急剧膨胀和增长。如何对 TB 级的大数据进行分析是一大挑战。

2）数据深度分析需求的增长

为了从数据中发现知识并加以利用进而指导人们的决策，必须对大数据进行深入的分析，而不是仅仅生成简单的报表。这些复杂的分析必须依赖于复杂的分析模型，很难用SQL来进行表达，统称为深度分析。人们不仅需要通过数据了解现在发生了什么，更需要利用数据对将要发生什么进行预测，以便在行动上做出一些主动的准备。

比如通过预测客户的流失预先采取行动，对客户进行挽留。这里，典型的OLAP数据分析操作（对数据进行聚集、汇总、切片和旋转等）已经不够用，还需要路径分析、时间序列分析、图分析、What-if分析以及由于硬件/软件限制而未曾尝试过的复杂统计分析模型等。

3）自动化、可视化分析需求的出现

因为数据规模很大，要对大数据进行有效分析，分析过程需要按照完全自动化的方式进行。这就要求计算机能够理解数据在结构上的差异，明白数据所要表达的语义，然后“机械”地进行分析。

对大数据分析来说，设计一个好的适于分析的数据表示模式是非常重要的。此外，大数据也使下一代可实时应答的交互式数据分析成为可能。将来，系统应该能够根据网站的内容自动构造查询，自动提供热门推荐，自动分析数据的价值并决定是否需要保存。目前，在保证交互式响应的同时如何进行TB级的复杂查询处理已成为一个重要的研究课题。

2. 主要进展

针对上面提到的挑战，研究者提出了一些试验性的解决方法和途径，其中的许多方法具有一定的实际应用价值。例如，针对传统分析软件扩展性差以及Hadoop分析功能薄弱的特点，IBM公司的研究人员致力于对R和Hadoop进行集成。R是开源的统计分析软件，通过R和Hadoop的深度集成，把计算推向数据并且并行处理，使Hadoop获得了强大的深度分析能力。另有研究者实现了Weka（类似于R的开源的机器学习和数据挖掘工具软件）和MapReduce的集成。

标准版Weka工具只能在单机上运行，并且不能超越1GB内存的限制。经过算法的并行化，在MapReduce集群上，Weka不仅突破了原有的可处理数据量的限制，轻松地对超过100GB的数据进行分析，同时利用并行计算提高了性能。经过改造的Weka，赋予了MapReduce技术深度分析的能力。

另有开发者发起了Apache Mahout项目的研究，该项目是基于Hadoop平台的大规模数据集上的机器学习和数据挖掘开源程序库，为应用开发者提供了丰富的数据分析功能。

针对频繁模式挖掘、分类和聚类等传统的数据挖掘任务，研究人员也提出了相应的大数据解决方案。如，Iris Miliaraki等人提出了一种可扩展的在MapReduce框架下进行频繁序列模式挖掘的算法，Alina Ene等人用MapReduce实现了大规模数据下的k-center和k-median聚类方法，Kai-Wei Chang等人提出了针对线性分类模型的大数据分类方法。U Kang等人使用Belief Propagation算法（简称BP）处理大规模图数据发掘异常

模式。

另有一些研究针对大规模图数据进行分析。Jayanta Mondal 等人提出了一个基于内存的分布式数据管理系统来管理大规模动态变化的图以支持低延迟的查询处理方法，提出了一种混合的复制(replication)策略来检测结点读写的频率从而动态的决定哪些数据需要复制(replication)。Shengqi Yang 等人对基于集群上的大规模图数据管理和局部图的访问特征(广度优先查询和随机游走等)进行研究，为了在图查询处理中减少机器间的通信，提出来分布式图数据环境，同时提出了两级别划分管理架构。Jiewen Huang 等人提出了一个多结点的可扩展 RDF 数据管理系统，比目前系统的效率高出三个数量级。

3. 发展趋势

1) 更加复杂、更大规模的分析和挖掘

在大数据新型计算模式上实现更加复杂和更大规模的分析和挖掘是大数据未来发展的必然趋势。例如，需要进行更细粒度的仿真、时间序列分析、大规模图分析和大规模社会计算等等。另一方面，在大数据上进行复杂的分析和挖掘，需要灵活的开发、调试、管理等工具的支持。

2) 大数据的实时分析和挖掘

面对大数据，分析和挖掘的效率成为此类大数据应用的巨大挑战。尽管可以利用大规模集群并行计算，以 MapReduce 为代表的并行计算模型并不适合高性能的处理结构化数据的复杂查询分析。在数十 TB 以上的数据规模上，分析和发掘的实时性受到了严峻的挑战，是目前尚未彻底解决的问题。而查询和分析的实时处理能力，对于人们及时获得决策信息，做出有效反应是非常关键的前提。

3) 大数据分析和挖掘的基准测试

各种大数据分析和挖掘系统各有所长，其在不同类型分析挖掘下，会表现出非常不同的性能差异。目前迫切需要通过基准测试，了解各种大数据分析和挖掘系统的优缺点，以明确能够有效支持大数据实时分析和挖掘的关键技术，从而有针对性地进行深入研究。

10.6　大数据可视化分析技术发展前景

1. 问题与挑战

在大数据时代，数据的数量和复杂度的提高带来了对数据探索、分析、理解和呈现的巨大挑战。除了直接的统计或者数据挖掘的方式，可视化通过交互式视觉表现的方式来帮助人们探索和解释复杂的数据。一个典型的可视化流程是首先将数据通过软件程序系统转化为用户可以观察分析的图像。

利用人类视觉系统高通量的特性，用户通过视觉系统，结合自己的背景知识，对可视化结果图像进行认知，从而理解和分析数据的内涵与特征。同时，用户还可以交互地改变可视化程序系统的设置，改变输出的可视化图像，获得对数据的不同侧面的理解。因此可视化是一个交互与循环往复的过程，如图 10.3 所示。

可视化能够迅速和有效地简化与提炼数据流，帮助用户交互筛选大量的数据，可视化

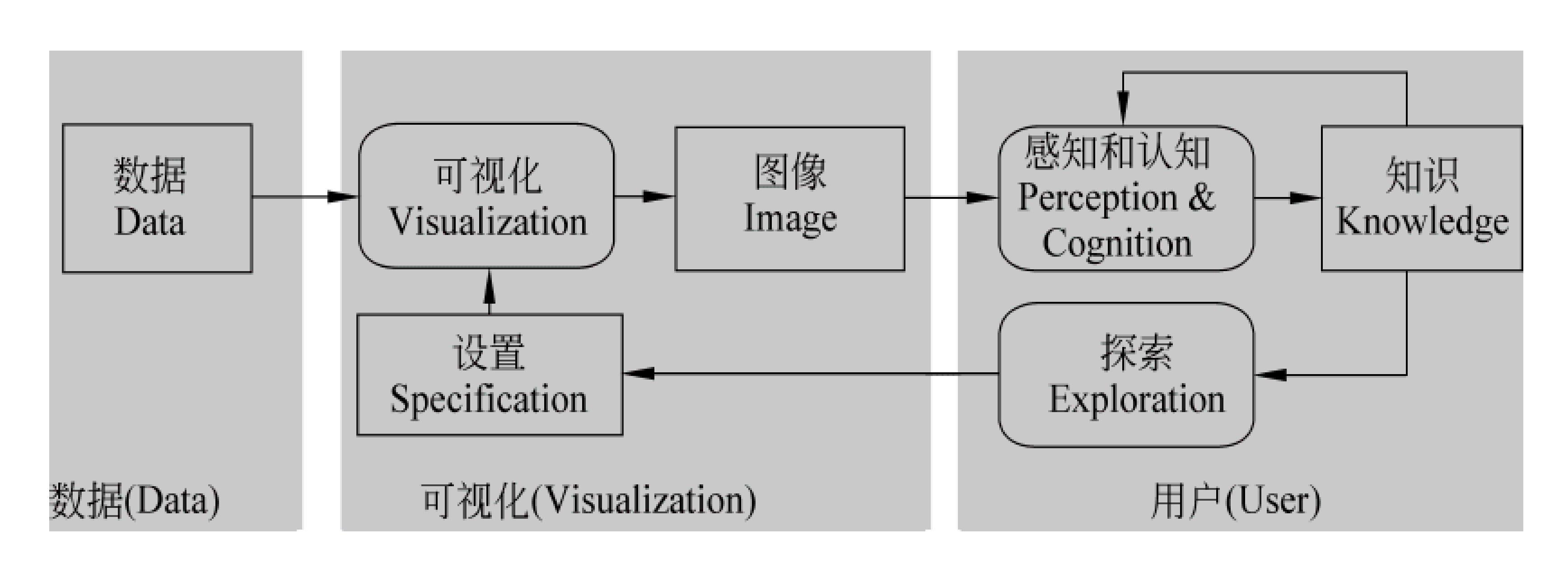


图 10.3 可视化流程

所提供的洞察力有助于使用者更快更好地从复杂数据中得到新的发现，这使得可视化成为数据科学中不可或缺的重要部分。人类对于数据对象通过作图的方式帮助理解分析古已有之。例如古人的地图和星图，早期物理学家对实验结果的绘图。现代意义上的可视化源自于计算机技术的发展，首先是对于科学数据的可视化，其后扩展到更广泛的信息可视化。进入 21 世纪后，随着反恐等需求，对于海量、复杂数据的分析进一步催生了可视分析，通过可视界面，结合人机交互和背景自动数据分析挖掘，对海量复杂数据开展分析。

2. 主要进展

在可视化的发展中，首先面对大规模数据挑战的是在科学可视化方向。高通量仪器设备、模拟计算以及互联网应用等都在快速产生着庞大的数据，对 TB 乃至 PB 量级数据的分析和可视化成为现实的挑战。大规模数据的可视化和绘制主要是基于并行算法设计的技术，合理利用有限的计算资源，高效地处理和分析特定数据集的特性。很多情况下，大规模数据可视化的技术通常会结合多分辨率表示等方法，以获得足够的互动性能。在科学大规模数据的并行可视化工作中，主要涉及数据流线化(Data Streaming)、任务并行化(Task Parallelism)、管道并行化(Pipeline Parallelism)和数据并行化(Data Parallelism)四种基本技术。

数据流线化将大数据分为相互独立的子块后依次处理。在数据规模远远大于计算资源时是主要的一类可视化手段。它能够处理任意大规模的数据，同时也可能提供更有效的缓存使用效率，并减少内存交换。但通常这类方法需要较长的处理时间，难以提供对数据的交互挖掘。离核渲染是数据流线化的一种重要形式。在另外一些情况下，数据则是以流的形式实时逐步获得，必须要有能够适应数据涌现形式的可视化方法。

任务并行化是把多个独立的任务模块平行处理。这类方法要求将一个算法分解为多个独立的子任务，并需要相应的多重计算资源。其并行程度主要受限于算法的可分解粒度以及计算资源中结点的数目。管道并行化则是同时处理各自面向不同数据子块的多个独立的任务模块。以上任务并行化和管道并行化两类方法，如何达到负载的平衡是实现高效分析的关键难点。

数据并行化是将数据分块后进行平行处理，通常称为单程序多数据流(SPMD)模式。这类方法能达到高度的平行化，并且在计算结点增加的时候可以获得较好的可扩展性。对于非常大规模的并行可视化，结点之间的通信往往是制约因素，提供合理的通信模式是高效结果的关键，而提高数据的本地性也可以大大提高效率。以上这些技术往往在实践

中相互结合，从而构建一个更高效的解决方法。

在信息可视化和可视分析方面，相对对大规模数据的处理，其出现的相应要晚得多。很多技术，例如多维数据可视化中的平行坐标、多尺度分析、散点图矩阵、层次数据可视化中的树图、图可视化中的多种布局算法、文本可视化的一些基本方法，并不是都有很好的可扩展性。在面对大数据挑战的可视化中，需要做出相应的调整。

传统对网络数据的可视化可以通过图的形式实现，这是将网络中的每个结点简化为图中的结点，网络中的联系可视化为图中的边，这样网络数据的可视化可以通过经典的结点-边的形式表现。这类可视化方法的难点主要在于图的排布算法。有效的图布局应该能够直观地揭示结点之间的联系，类似地，相互联系紧密的结点会聚集在一起。但是现在大规模的网络数据的结点可能高达数百万，其边可能高达数亿，这样的网络数据难以使用传统的图可视化方法可视化。

高维信息可以通过维度压缩、平行坐标等手段实现可视化。但是在数据达到一定规模以后，这样的方法并不能很好地扩展。一些可能的方案包括提供一些子空间的选择，用户可以根据分析需要，在高维度空间选择适合问题解决的子空间，从而缩小数据规模。

图形硬件对于大规模数据可视化具有重要意义。最新的超级计算机大量地应用 GPU 作为计算单元。如何更好地发掘最新的图形硬件潜力，提供更加灵活的大数据可视化和绘制的解决方法是具有重大意义的课题。

3. 发展趋势

面对大数据，结合国际学者的各种观点，相应的大数据可视化与分析也面临着各种挑战。

1）原位分析(In Situ Analysis)

传统的可视化方式是先将数据存储于磁盘，然后根据可视化的需要进行读取分析。这一种处理方式对于超过一定量级的数据来说并不适合。最初是为了应对超大规模的超级计算机计算获得的大量科学数据产生的挑战。科学家提出了原位可视分析的概念，在数据仍在内存中时就会做尽可能多的分析。

对数据进行一定的可视化(同时也是数据规模的简化)，能极大地减少 I/O 的开销，只有极少数的视觉投影后的次生数据需要转移到显示平台。这个方法可以实现数据使用与磁盘读取比例的最大化，从而最大限度地克服 I/O 的瓶颈限制。

然而，它也带来了一系列设计与实现上的挑战，包括交互分析、算法、内存、I/O、工作流和线程的相关问题。原位分析要求可视化方案和计算紧密结合，这样很多传统的可视化方法需要进行修改或者筛选才可以用于这样的可视化模式。由于可视化的一部分处理在计算核点上进行，那样就会对可以进行的处理方案有所限制。

2）大数据可视化中的人机交互

在可视化和可视分析中用户界面与交互设计扮演着越来越重要的角色。用户必须通过合理的交互方式，才可以有效地探索发现数据中的隐含信息，进行可视推理，通过意义构建，获得新的认知。然而尽管数据规模和机器的计算能力都在持续快速地增长，千百年

来,人的认知能力却是始终不变的。以人为中心的用户界面与交互设计面临的挑战是复杂和多层次的,并且在不同领域都有交叠。

机器自动处理系统对于一些需要人类参与判断的分析过程往往表现不佳。其他的挑战则源于人的认知能力,现有技术不足以让人的认知能力发挥到极限。我们需要提供更好的人机交互界面和设计,方便使用者,特别是专家用户能够最大程度地发挥其背景知识,在数据的分析中扮演更加积极的角色。从更广泛的意义上说,可视化可以建立一个可视的交互界面,提供人和数据的对话。

3)协同与众包可视分析

在大数据时代,个人或者少数几个分析用户可能无法面对数据规模和复杂度带来的挑战。大数据分析中往往会设计多种不同来源甚至领域的数据。利用众人的智慧,通过众包等模式进行有效的复杂可视化成为一种必然的选择。在众包可视化工作中,如何设计合理高效的可视化平台,承载相应的复杂高难度的可视化系统工作;如何设计交互的中间模式,支持多用户的协调工作;如何反映多用户的差别,都是可以研究的课题。和协同的可视分析方式比较,协同可视化趋于少数的几个领域专家交互合作开展对数据的可视分析,众包可视化则更趋向不特定多数的使用者,规模也更大。如何开展有效的众包和协同可视化,是非常重要的研究课题。

4)可扩展性与多级层次问题

在大规模数据可视分析的可扩展性问题上,建立多级层次是主流的解决办法。这种方法可以通过建立不同大小的层面,向用户提供在不用解析度下的数据浏览分析能力。但是当数据量增大时,层级的深度与复杂性也随之增大。在继承关系复杂且深度大的层次关系中巡游与搜索最优解是可扩展性分析的主要挑战。

5)不确定性分析和敏感性分析

不确定性的量化问题可以追溯到由实验测量产生数据的时代。如今,如何量化不确定性已经成为许多领域的重要问题。了解数据中不确定性的来源对于决策和风险分析十分重要。随着数据规模增大,直接处理整个数据集的能力也受到了极大的限制。许多数据分析任务中引入数据的不确定性。不确定性的量化及可视化对未来的大数据可视分析工具而言极端重要,我们必须发展可应对不完整数据的分析方法,许多现有算法必须重新设计,进而考虑数据的分布情况。

一些新兴的可视化技术会提供一个不确定性的直观视图,来帮助用户了解风险,从而帮助用户选择正确的参数,减少产生误导性结果的可能。从这个方面来看,不确定性的量化与可视化将成为绝大多数可视分析任务的核心部分。

另一方面,对于可视化而言,用户的交互或者新的参数的输入,都会导致不同可视化结果的出现。在大数据的情况下,向用户提供背景知识,告知预期的操作可能引发的可视化结果的变化程度,或者用户当前所在参数空间的周边状况,这一些都属于对可视分析结果的敏感性分析,对于高效的可视化交互是极端重要的。

6)可视化与自动数据计算挖掘的结合

可视化提供了用户对数据的直观分析,用户可以通过交互界面对数据进行分析了解。同时,我们要注意到很多的数据分析是批量的。如何能够将一些比较确定的分析任务利

用机器自动完成,同时引导用户来进行更具有挑战性的可视分析工作,是可视分析发展中的核心课题。

7）面向领域和大众的可视化工具库

提供相应的工具库可以大大提高不同领域分析数据的能力。大数据时代涌现并推动了很多可视化商业化的机会。Tabealu 的成功上市反映了市场对可视化工具的需求。类似 IBM Manyeyes 这样在线可视化工具的流行,则表明在一定程度上满足了广大普通用户对可视化方法的需求。国际的几个大公司也在开展相应的研究,企图把可视化引入其不同的数据分析和展示的产品中。

各种可能相关的商品也将会不断出现,对可视化服务的商业需求将是未来的一个最大方向。

10.7　大数据隐私与安全技术发展前景

1. 问题与挑战

隐私是当事人不愿意被他人知道或他人不便知道的敏感信息,它与公共利益、群体利益无关,具有隐藏特性。安全是指不受威胁,没有危险、危害、损失。信息安全是指采取技术和管理的安全保护手段,保护软硬件与数据不因偶然的或恶意的原因而遭到破坏、更改、显露。

在大数据时代,传统的隐私数据内涵与外延有了巨大突破与延伸,隐私数据保护不力所造成的恐慌已不能由个人或团体承受,隐私数据保护技术面临更多的挑战。大数据时代下的隐私数据保护与安全体系除涉及技术、管理外,还涉及法律、人伦、生物、道德、商业利益、生活方式等;不只是团体或区域,还涉及国家安全与国际秩序。隐私数据泄露影响的波及面很可能会突破个人、团体或区域的限制,发展到全球性影响。

从本质上来说,大数据的安全与隐私问题就是我们要能够在大数据时代兼顾安全与自由,个性化服务与商业利益,国家安全与个人隐私的基础上,从数据中挖掘其潜在的巨大商业价值和学术价值,并使其研究成果真正地服务于社会。

在大数据时代,随着人们对大数据的进一步认识和研究,呈现出的安全隐私挑战体现几个方面：

(1) 大数据时代的安全与传统安全相比,变得更加复杂。

一方面,大量的数据汇集,包括大量的企业运营数据、客户信息、个人的隐私和各种行为的细节记录。这些数据的集中存储增加了数据泄露风险,而这些数据不被滥用,也成为人身安全的一部分。另一方面,大数据对数据完整性、可用性和秘密性带来挑战,在防止数据丢失、被盗取和被破坏上存在一定的技术难度,传统的安全工具不再像以前那么有用。

(2) 使用数据过程中的安全问题。

用数据挖掘和数据分析获取商业价值的时候,黑客也可以利用大数据分析向企业发起攻击。黑客可能会最大限度地收集有用信息,如社交网络、邮件、微博、电子商务、电话和家庭住址等,使得数据安全局面异常严峻。

(3) 对大数据分析较高的企业和团体,面临更多的安全挑战。

对于电子商务、金融、天气预报的分析预测、复杂网络计算和广域网感知等领域,恶意攻击会造成更会严重的后果。

(4) 基于位置的隐私数据暴露严重。

随着个体用户的移动设备,如手机、移动 GPS 设备等的广泛使用,以及通过一些网站获取用户位置信息等可以很容易得到用户的移动轨迹。而根据研究发现,用户的移动模式和用户身份识别之间有着强烈的对应关系,使得用户的隐私很容易暴露。同时,用户的位置信息保护比用户的身份信息保护更具有挑战性,因为我们在获取数据时要保证较高的精度。

(5) 缺乏相关的法律法规保证。

目前为止,还没有严格的法律法规来保证用户的数据隐私安全。特别是一些涉及用户敏感数据的一些记录,而这些数据也容易被一些非法和不道德组织或个体使用,对用户和社会造成严重的影响和损失,例如,频繁发生的互联网公司数据库泄露事件,特别是2013 年曝光的美国国家安全局“棱镜计划”监听项目。

(6) 大数据的共享问题。

共享问题的主要本质是数据的加密性和数据的有效性之间的矛盾。从社会应用角度考虑,我们会尽可能提高数据的获取技术,以保证数据的有效性,而从保护用户隐私的角度考虑,我们有必要对数据进行相关操作以降低获取数据的敏感性,从而造成了两者之间的矛盾,两者之间如何进行最佳折中确实非常困难。

(7) 真实数据的动态性变化。

具有真实性的大数据随着时间呈现出动态变化性,使得我们对于大数据的分析计算提出了一些新的方法和技术,因而在处理时将面对更为复杂的形式,加大了大数据安全隐私保护的困难。

(8) 多元数据的融合挑战。

大数据来自于生活、学术、商业等各个方面,而数据之间的彼此相关性,使得数据的安全隐私保护更为复杂,如何在多元数据融合的大趋势下保证用户的隐私不被泄露是一项重大挑战。

2. 主要进展

数据的安全与隐私问题近年来一直是国内外学者关注的重大研究课题,并且针对不同的应用和数据类型都有相关的研究成果,总的来说,目前所使用的方法有:

(1) 文件访问控制技术。

通过文件访问控制来限制呈现对数据的操作,在一定程度解决数据安全问题。

(2) 基础设备加密。

其本质是对大数据的存储设备进行安全防护,但不能解决大数据安全的本质问题。

(3) 匿名化保护技术。

匿名化技术适用于各类数据和众多应用,并且算法通用性高,能保证发布数据的真实性,实现简单。匿名化过程不可逆,如决策分类器的构建、聚类等应用,如 k-匿名模型,m-

invariance 等。但匿名化技术对隐私保护效果并不明显，使得隐私泄露可能性很大。

（4）加密保护技术。

加密保护技术能够保证数据的真实性、可逆性和无损性，对隐私保护程度很高，主要应用于分布式下的数据挖掘和操作，如 SMC 模型、分布式关联规则挖掘算法、差分隐私等。但是该技术的计算开销很大，对大数据的支持不大适用。

（5）基于数据失真的技术。

该技术可应用与关联规则的挖掘和隐藏等，如随机干扰、随机化、阻塞、凝聚等。数据失真技术的实现比较简单，但会造成数据的偏差，可能造成数据价值的丧失。

（6）基于可逆的置换算法。

可逆的置换算法可以保证数据的真实性，并且效率比较高，常用于数据中心的大规模系统隐私保护，如位置变换、映射变化等，但该技术对于安全隐私保护力度仍然不够充分。

3. 发展趋势

随着大数据的不断发展和研究，其巨大价值在被不断挖掘的过程中，数据的安全和隐私发展呈现出新的发展趋势和挑战。

（1）NoSQL 有待进一步完善：迎合了大数据的时代，适合非结构化数据的存储和分析，有灵活、可扩展性强、降低复杂性等特点，但是在安全保护上有待进一步提高。

（2）针对 APT 的攻击：在大数据时代，我们在利用数据来获取价值，APT 的攻击隐藏在数据内部，很难被我们发现，所以专门针对 APT 攻击的研究是非常重要的。

（3）大数据的迅速发展和数据量的急剧增加及急速的动态变化，使得我们在对数据的操作时所面临的安全问题更加严重。

（4）数据的多元化与彼此的关联性进一步发展，深度挖掘技术、分析方法、算法模型的进一步优化和提高，使得对单一数据的安全隐私保护方法变得极其脆弱，需要针对多元数据融合提出新的安全隐私保护技术。

（5）针对目前的大数据计算，主要采取的是分布式计算方法。而采用分布式计算的时候必然面临着数据传输，信息交互等过程，如何在这个过程中保护数据不泄露、信息不丢失、保护所有站点的安全与分布式系统的隐私是大数据发展面对的重大挑战。

（6）目前，社交网络成为现代生活不可或缺的部分，一般来说，社交网络都会获取个体用户的位置信息（如 Facebook、新浪微博等），基于网络的迅速动态变化和实时交互等性质，使得我们对网络的安全加密与数据保护更为困难，而作为目前迅速发展起来的社交网络，我们需要进一步加强此方面的安全隐私保护。

（7）“三权分立”的模式应成为一种趋势，即数据的采集过程保护、存储管理保护以及数据的分析使用过程的安全保护需要由不同的管理决策者来执行，这样可以在一定程度上保护大数据的安全隐私。

最后，大数据的保护需要学术界、商业界以及政府部门的共同参与，需要形成有效的安全机制和国家法律法规来约束和保护大数据的安全隐私，从而保证大数据时代的健全、安全发展。

10.8 大数据应用案例之：数据解读城市：北京本地人VS外地人

在各大城市，“外来人口”都是一个随时可以引起争议的话题。

毫无疑问，外来者为城市的经济发展、城市运行注入活力，同时也是公共资源的使用者。在人口、资源、环境等压力之下，城市的拥挤、无序、污染、不文明等往往被当地居民归咎于“外地人”。

互联网自媒体和各个地方论坛上，“本地人”对“外地人”的抱怨和“外地人”对“本地人”的反击非常普遍。在这些争论中，我们时常见到本地人指责外地人对原有城市生活环境、文化、语言等方面造成冲击。

那么，首都的情况如何？近年来，随着北京城市面积的扩张和旧城人口的疏解，北京的本地论坛里也出现“北京首都化的过程就是外地人进三环，北京土著出五环的过程”这一说法。这些观感究竟是带着情绪的抱怨，还是某种程度上的事实？我们无意介入具体的争论，这里仅基于数据，从常住外来人口、外来青年人才以及短期来京外来人口三个视角，分析一下北京外来人口的分布情况。

大数据论证：你的上班路为何会变成漫长“取经”路？（北京）

北京的人流在哪儿？用大数据看城市。

用数据来勾画，24:00之后的北京到底是啥样儿？

大数据颠覆您心中的房奴形象(来自2012—2014年的50000+北京商业贷款案例)。

视角一：常住外来人口分布

北京的“外地人”，到底住在哪些地方？是不是真的把原先的老北京“挤出去”了呢？我们先来试着回答一下这个问题。

从2010年全国第六次人口普查的数据看，北京常住外来人口数量的分布图如图10.4所示。

常住人口指的是，“全年经常在家或在家居住6个月以上，且经济和生活与本户连成一体的人口”。北京市常住外来人口总共有702.8万，其中一半以上居住在五环之外，具体数字是375.2万。

我们以“乡镇街道办事处”为统计单位，考察常住外来人口总数，发现大量外来人口集中在五环以外(颜色越深，常住外来人口数量越多)：北五环至六环之间的回龙观、东小口、北七家；南三环与南五环之间卢沟桥、新村、大红门旧宫十八里店；以及东五环外平房和永顺街道，如图10.5所示。

通过“外来人口所占单元总人口比例”数据观察本地外来人口分布，我们发现，在各环路之间，五六环之间的常住外来人口占全部常住人口的比例最高，达到61.8%，而四环内常住外来人口占比仅约为32.1%。

小结：

与上海所呈现的现象不同的是，北京本地人(拥有北京户籍的人口)，还是比较住在繁

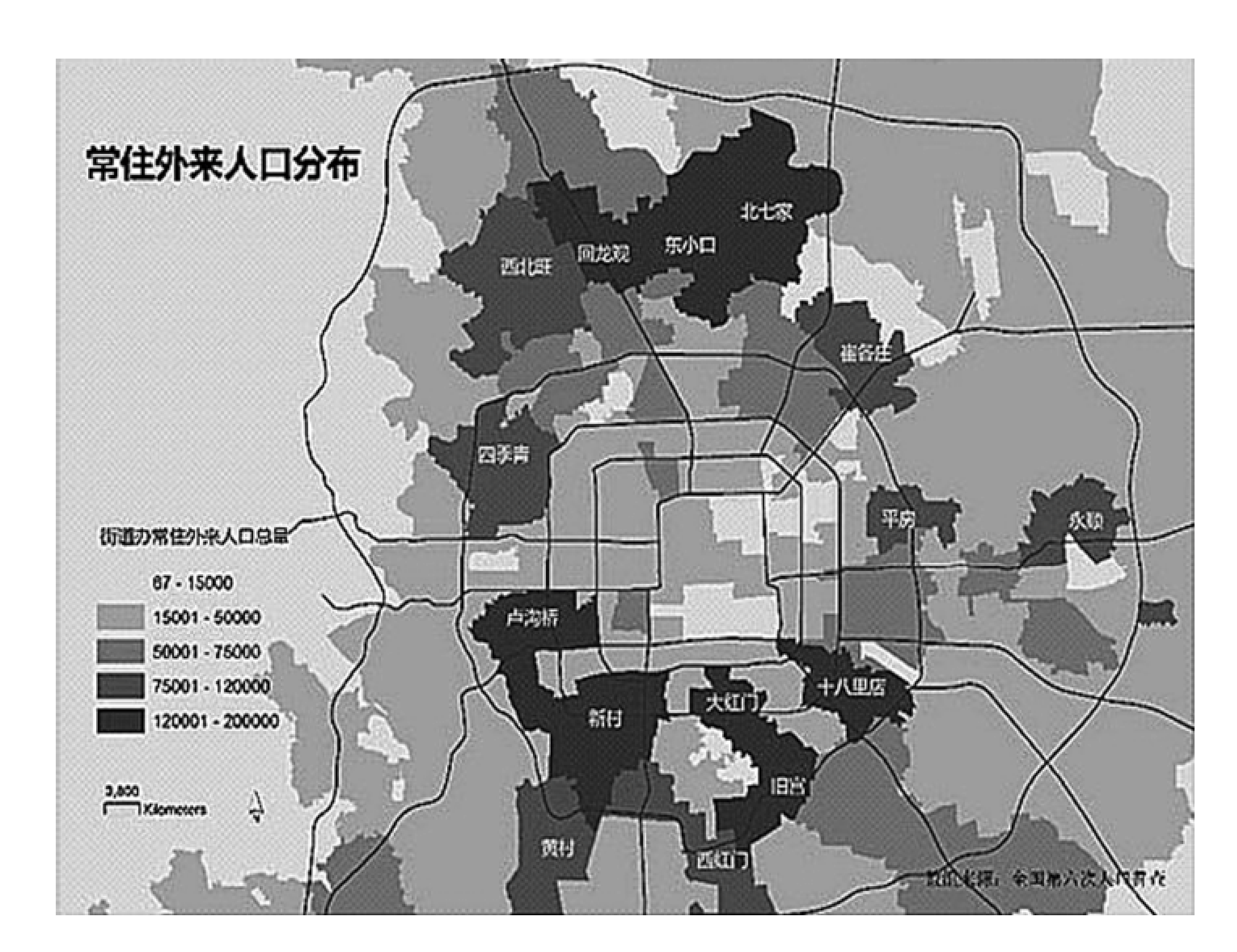


图 10.4　北京常住外来人口分布

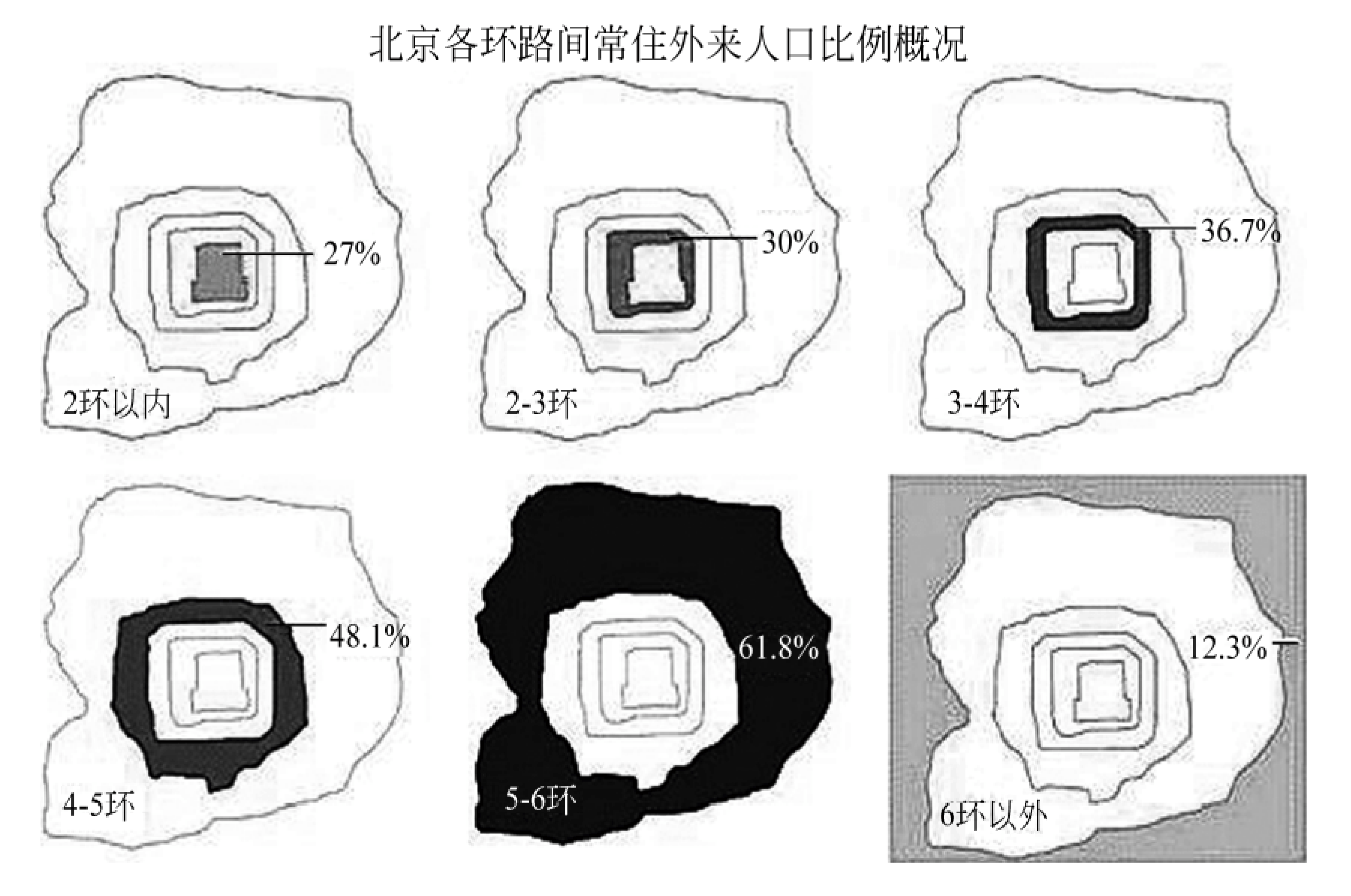


图 10.5　北京各环路间常住外来人口比例概况

华的“城里”，尽管存在人口疏解措施，但截至2010年，四环以内北京本地人为主的人口结构，并没有改变。

视角二：“外来人才”分布

对于特大城市的政府而言，在对外来人口限制的同时，对所谓“外来高端人才”通常持欢迎态度。那么，是否意味着，“外来高端人才”会住在北京靠近市中心一带？

由于缺乏“外来高端人才”的官方统计口径，为了观察他们的分布，我们用“拥有大学及以上学历的青年人”来分析。根据某电商的用户画像数据，我们分析了20～30岁大学及以上学历的工作人群在京的分布情况。

总体上看，没有本地人才绝对主导的区域，却有一些居住单元或就业单元，青年人才里90%以上是外地人。

我们发现，外来青年人才大量安居(或租住)在海淀东部北部、朝阳、顺义、通州、亦庄、大兴等区域，在城市北、东、南三个方位形成一个倒C的包围圈。

尤其是在回龙观、天通苑、沙河、宋庄以及黄村等部分区域，如果你遇到一个有大学以上学历的青年人，那么他/她有90%以上的可能会是个外地人。

在北京，外来青年人才平均通勤距离近20km，他们从事信息技术、软件、互联网、新材料、新型制造业等高新技术行业工作。而本地年轻人才更多居住在东城、西城、海淀西南部、丰台东南和河西、门头沟、房山等区域，呈带状分布。

本地青年人才居住比例(即青年人才中，本地人比例)最高的单元，是长辛店的63%和苹果园的55%；在中心城区，本地青年人才居住比例最高的单元依次是新街口55%，右安门54%，景山52%，以及大栅栏50%。就业方面，本地青年人才比例最高的单元依然是长辛店的56%和苹果园的55%，三环内中心区域本地人才就业比例较高的单元有交道口54%、大栅栏53%、白纸坊52%、东铁匠营52%和展览路51%。

大体上看，在国家部委、机关事业单位、文化、医疗、商贸等岗位密集的区域，本地人才比例稍高；在金融、教育科研、文创产业等行业密集的区域，本地与外地人才的比例相似。

可见，尽管政府欢迎“外来高端人才”，但“人才”中的大部分却并未进到城里。

说明：本视角观察的“本地人”和“外来人”并非是城市的全部人口结构，而是20～30岁已毕业的具有大学及以上学历的群体。此外，与常住外来人口的户籍区分方式不同，此处的“本地”“外地”主要按照出生地来分辨。

视角三：短期来京者的分布

以上两个视角观察的都是定居北京或在京长期就业的外来人口分布，至此，我们并未发现显著的“外来人口主导北京城”的现象，尤其是，在东城区和西城区的常住人口和就业人群里，北京本地人占绝大多数。

但是，为什么在很多人印象中，北京城里四处都是操着异地口音的“外地人”？实际上，北京城的活动人群中，有大量短期外来人员，比如游客、探亲访友者、来京出差的商务人士等。他们并没有出现在我们前面的统计之中。

为了观察这类人群的分布特征，我们利用“人迹地图”大数据平台，基于2015年某普通工作日定位数据，对北京东城和西城区，以及短期外来人员较多的热门吸引点，识别了常住地在北京的人和常住地不在北京的短期来京者，观察他们一日内(上午、下午、夜间)在北京的分布情况。这些热门地点可分为办公、商业、景点、对外枢纽和批发市场五大类。

从整体上看，排除短期外来人员集中的机场、火车站等对外交通枢纽，白天的国贸区域是短期外来人员最密集的区域，共观测到近3万人，密度约0.8万人/km^2。国贸区域是北京的商务中心区，在工作日，有大量出差来京的商务人士前往。此外，同时这里也是重要的公共交通结点。据报道，国贸地铁换乘日均人流量可达30万人次，故该区域观测到的人口数据一定程度上受到了地铁、公交客流影响。

八达岭全天观测到1.6万短期外来人员，不过由于占地面积广阔，八达岭长城的人员

密度并不高；天安门区域虽然观测到的短期外来人员绝对数量不算靠前，却是短期外来人员密度最高的区域；王府井则在数量和密度上都位居前列，如图 10.6 所示。

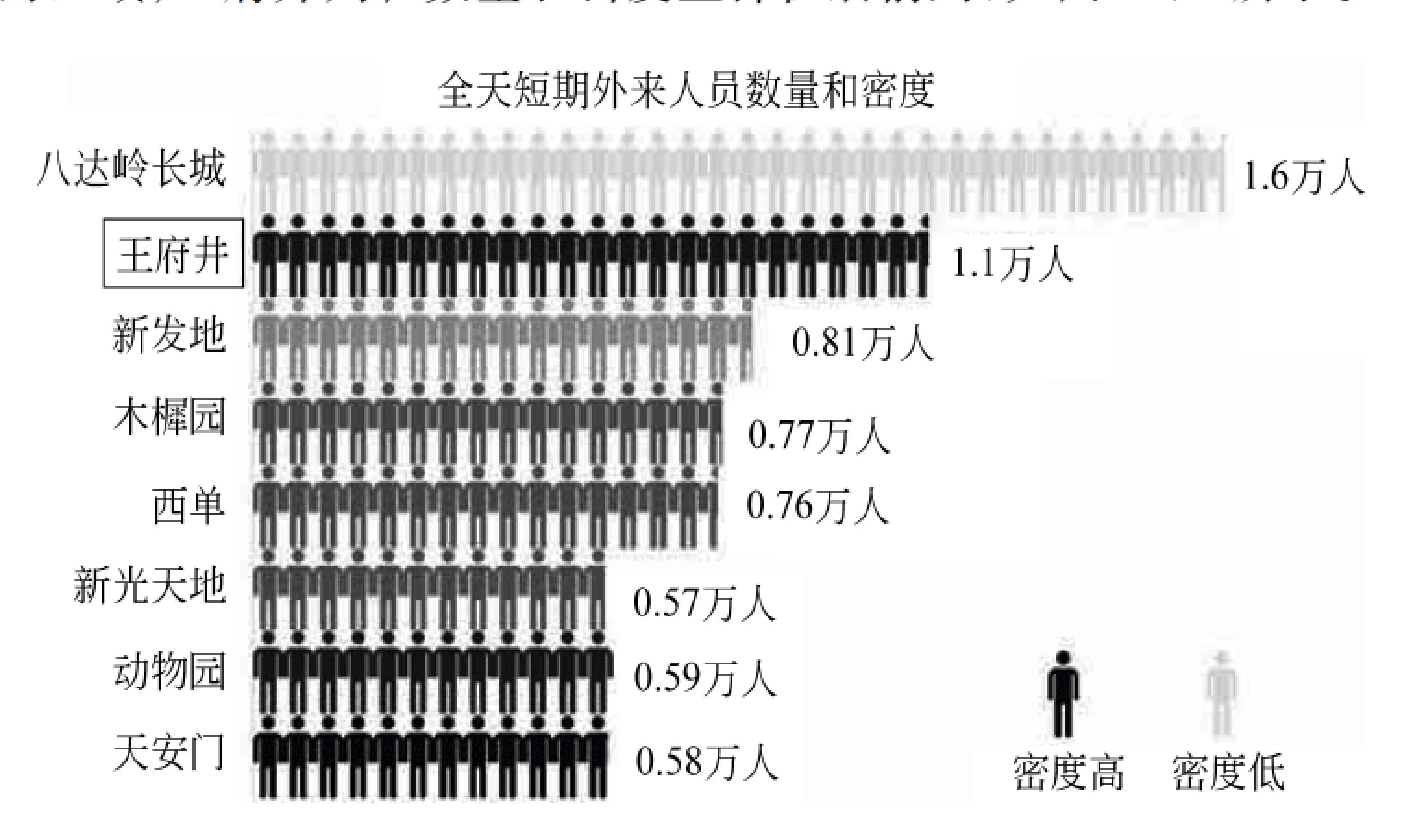


图 10.6　代表性吸引点全天短期外来人员观测数量和密度

进一步按照上午、下午和晚上对各类吸引点单元做详细分析。

1. 东城，西城

从东城区、西城区全区看，在该工作日的上午，东城区总共观测到约 127 万人，其中短期外来人员约占 16.7％，西城区总共观测到约 140 万人，短期外来人员约占 18.9％；下午与夜间的比例相似。也就是说，工作日在北京旧城出现的人当中，每五个人就有一个是短期外来人员。

2. 办公类

国贸区域上午观测到人口 21 万人，下午则增加至 23 万人。如果你这个工作日白天出行目的地是国贸，那么在这边碰到的人里，有 13％的概率是短期外来人员；到了夜间，短期来京者比例增加至 23％，这是因为在京定居的商务精英下班离开 CBD，而不少外来的商旅人士则选择在 CBD 区域的商务酒店里度过这个普通的夜晚。

3. 景点类

天安门向来是外地游客必去的景点，在我们观测的 33 个热门吸引点中，这里是唯一出现过短期外来人员比例高于本地人的地点。在这个工作日的上午，这里观测到 1.4 万人，其中有 52％是短期来京者。到了下午，在天安门区域的可识别人数减少了 4 千人，短期来京者的比例也下降到了 38％。天安门是游客游览故宫的入口，大多数游客会选择早上进入故宫。此外，这里早上有万众瞩目的升国旗仪式，毛主席纪念堂也只在上午开放。所以，如果你这天早上来到天安门，那么你遇到的路人是外地游客的概率会大于是北京市民的概率。

跟天安门相反，南锣鼓巷下午更吸引游客，且更吸引本地游客，短期来京者比例并不高。该日上午共观测到 2 万人，下午增至 2.5 万人，短期来京者的比例也从 14％增至 15％，到了夜间，短期来京者比例进一步增至 17％。

颐和园占地面积较大，游客往往要花费一整天时间才能走完。因此上午和下午可观

测的人数浮动在1.6万人上下，并没有太大变化。其中短期外来人员比例也很高：在与你擦肩而过的游人中，有40%的可能是外地游客。

而前面提到的八达岭长城，白天观测到4.2万人，外地游客比例同样在40%左右。

4. 商业类

西单商圈作为老牌商业中心，同样是吸引短期来京者的重要地点。在该工作日的上午，我们在西单观测到3.4万人，下午增至将近4.8万人，其中短期来京者的比例保持在18%左右。到了夜间，随着北京市民离开西单回家，短期来京者的比例增至28%，可以推测，很多来京出差、旅游的人，会选择住在生活和交通都很便利的西单附近。

王府井共观测到4万人，上下午总人数基本没有差异，其中有40%是短期来京者，这一比例远高于西单。到了夜间，依然有30%的短期来京者住宿在这里。由此看来，同样是国家级商业中心，王府井对外地人的吸引力比西单更高，后者还是以服务本地市民为主。

5. 对外枢纽类

在对外交通枢纽中，首都机场、北京站的短期外来人员比例在30%～35%，北京西站略高，约在45%左右；而北京南站的情况则比较特殊：我们发现，这里的短期来京者比例非常低，全天都维持在15%，与其他火车站以及首都机场相比相差甚大。这是否说明，北京南站这个价格稍高的高铁站，更多是为进出北京的北京市民服务？不过，北京南站我们观测到的人口数量相比起其全天吞吐量而言样本量太少，未来我们会寻找其他数据源来验证这个结论，如图10.7所示。

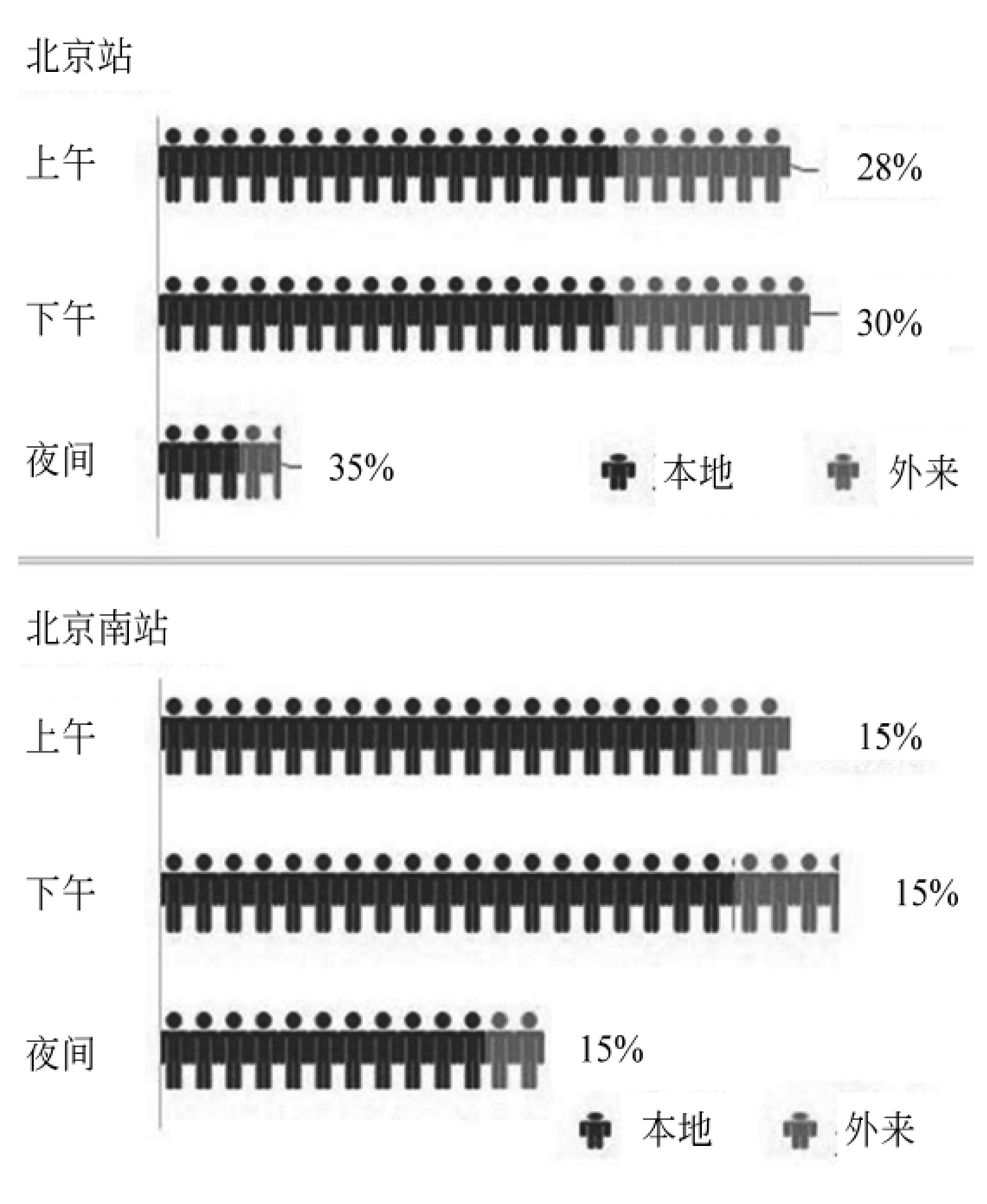


图10.7　对外枢纽2：北京站和北京南站对比

6. 批发市场类

在很多人印象里，批发市场是外地人密集的区域。动物园批发市场区域在上午观测到近 2.9 万人，下午增至 3.2 万余人。这里短期来京者的比例并不高，而且从早到晚比例变化都不大，仅在 20%左右。

位于西南四环外的新发地农产品批发市场白天观测到 3 万人左右，短期来京者约占 1/4，比例高于动物园批发市场；到了夜间，短期来京者的比例则增至 29%。因为从晚上 11 点到次日清晨 6 点之间，北京才允许外地车辆进城，新发地作为农产品批发市场，每天都有很多外地车辆在夜间送货到这里。

从数据来看，这些批发市场服务的对象仍是北京市民——尽管我们目前还难以分辨这里面有多大比例是北京户籍。

总结

经过以上分析，"北京首都化的过程就是外地人进三环，北京土著出五环的过程"这个说法并不正确。无论常住外来人口还是"外来人才"，他们都主要集中在城市北、东、南三个区域的四环、五环外围，四环以内仍以本地人为主，"外来人才"从事的行业与城里的本地人才从事的行业也开始产生了一定分化。可见北京的新城和边缘集团建设还是明显起到了对外来人口的截留作用——当然，城里高昂的房价和房租的作用也不容忽视。

但在二环以内、商务中心区以及主要旅游景点，有相当比例的人是短期外来人员，基本在 20%以上，在某些时刻某些区域甚至超过 50%。这些短期来京者，使北京市民产生某种错觉——北京城里外来人口太多。

近来，北京开始疏解城六区的"非首都功能"，但在城里的居住者和就业者主体上还是北京本地人。而旅游景点、商业中心和商务中心区是搬不走的。那么，当北京本地市民随着他们的安置房以及岗位而被疏解到新城和北京周边，那些留在北京城里的人们就会有更高的概率与来自祖国各地的同胞擦肩而过，届时他们或许会想——是不是疏解力度还不够？

参 考 文 献

1. 维克托·迈尔-舍恩伯格(Viktor Mayer-Schönberger),肯尼思·库克耶(Kenneth Cukier). 大数据时代:生活、工作与思维的大变革. 杭州:浙江人民出版社,2012.
2. 董西成. Hadoop 技术内幕:深入解析 MapReduce 架构设计与实现原理. 北京:机械工业出版社,2013.
3. 王星. 大数据分析:方法与应用. 北京:清华大学出版社,2013.
4. 赵刚. 大数据:技术与应用实践指南. 北京:电子工业出版社,2013.
5. 比约·布劳卿(Bjorn Bloching),拉斯·拉克(Lars Luck),托马斯·拉姆什(Thomas Ramge),大数据变革:让客户数据驱动利润奔跑. 沈浩译. 北京:清华大学出版社,2013.
6. 杨巨龙. 大数据技术全解:基础、设计、开发与实践. 北京:电子工业出版社,2014.
7. 埃里克·西格尔. 大数据预测. 周昕译. 北京:中信出版社,2014.
8. 赵伟. 大数据在中国. 北京:清华大学出版社,2014.
9. Thomas Erl,Zaigham Mahmood,Ricardo Puttini. 云计算:概念、技术与架构. 龚奕利译. 北京:机械工业出版社,2014.
10. 李军. 大数据:从海量到精准. 北京:清华大学出版社,2014.
11. 伊恩·艾瑞斯(Ian Ayres),大数据思维与决策. 宫相真译. 北京:人民邮电出版社,2014.
12. 西蒙(Phil Simon). 大数据应用:商业案例实践. 邓煜熙,漆晨曦,张淑芳译. 北京:人民邮电出版社,2014.
13. 陈明. 大数据概论. 北京:科学出版社,2014.
14. 赵勇. 架构大数据:大数据技术及算法解析. 北京:电子工业出版社,2015.
15. 莱斯科夫(Jure Leskovec),拉贾拉曼(Anand Rajaraman). 大数据:互联网大规模数据挖掘与分布式处理. 2 版. 北京:人民邮电出版社,2015.
16. Phil Simon. 大数据可视化:重构智慧社会. 漆晨曦译. 北京:机械工业出版社,2015.
17. 刘鹏. 云计算. 第三版. 北京:电子工业出版社,2015.
18. 王宏志. 大数据算法. 北京:清华大学出版社,2015.
19. 李俊杰,石慧,谢志明,谢高辉,唐华,王鹏云. 计算和大数据技术实战. 北京:人民邮电出版社,2015.
20. 林子雨. 大数据技术原理与应用:概念、存储、处理、分析与应用. 北京:人民邮电出版社,2015.
21. 林伟伟,刘波. 分布式计算、云计算与大数据. 北京:机械工业出版社,2015.
22. 安俊秀,王鹏,靳宇倡. Hadoop 大数据处理技术基础与实践. 北京:人民邮电出版社,2015.
23. 中国计算机学会大数据专家委员会. 中国大数据技术与产业发展白皮书(2013). 中国计算机学会,2013.
24. 中国电子信息产业发展研究院. 2015 大数据白皮书. 中国电子信息产业发展研究院,2015.